Uwe W. Gehring ·

Grundkurs
für Politolo
Soziologen

5., überarbeitete

Uwe W. Gehring · Cornelia Weins

Grundkurs Statistik für Politologen und Soziologen

Uwe W. Gehring · Cornelia Weins

Grundkurs Statistik für Politologen und Soziologen

5., überarbeitete Auflage



Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.

1. Auflage 1999
2. Auflage 2000
3. Auflage 2002
4. Auflage 2004
5., überarbeitete Auflage 2009

Lektorat: Frank Schindler

VS Verlag für Sozialwissenschaften ist Teil der Fachverlagsgruppe
Springer Science+Business Media.
www.vs-verlag.de



Umschlaggestaltung: KünkelLopka Medienentwicklung, Heidelberg
Druck und buchbinderische Verarbeitung: Krips b.v., Meppel
Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier
Printed in the Netherlands

ISBN 978-3-531-16269-0



Für Willy H. Eirmbter

Vorwort zur 5. Auflage

Die empirisch ausgerichteten Sozialwissenschaften verlangen von ihren Absolventen einen sicheren Umgang mit den Methoden der Datenerhebung und Datenanalyse. Auch im Studium der *Politikwissenschaft* und der *Soziologie* spielt die Beschäftigung mit den Techniken der empirischen Sozialforschung und der Statistik eine wichtige Rolle. Die Anwendung statistischer Methoden wurde sicher nicht zuletzt durch die rasante Entwicklung leistungsfähiger Personalcomputer und einfach bedienbarer Statistikprogramme begünstigt. Grundlegende Statistikkenntnisse sind jedoch nicht nur bei eigenen Analysen unabdingbar. Ein erheblicher Teil der sozialwissenschaftlichen Literatur kann ohne diese Kenntnisse nicht mehr nachvollzogen werden, wie man, um nur zwei Beispiele zu nennen, an den Artikeln in der *Politischen Vierteljahresschrift* oder der *Kölner Zeitschrift für Soziologie und Sozialpsychologie* nachvollziehen kann. Nicht zuletzt stellen Statistikkenntnisse – und die Beherrschung entsprechender Software – eine Schlüsselqualifikation für den Arbeitsmarkt dar.

Das Buch ist aus einem Manuskript entstanden, das wir für die Teilnehmerinnen und Teilnehmer unserer Kurse „Einführung in die Methoden der empirischen Sozialforschung und Statistik“ verfasst haben. Vom Umfang und Inhalt her ist das Buch für einen Kurs von vier Semesterwochenstunden konzipiert. Ziel ist die Vermittlung grundlegender Kenntnisse in den Methoden der Datenerhebung und der Statistik, die eine eigenständige Beschäftigung mit weiterführenden Methoden ermöglicht. Mathematische Vorkenntnisse werden nicht vorausgesetzt.

Das Buch gliedert sich in die Teile *Methodenlehre* (Kapitel 1 – 4), *Deskriptive Statistik* (Kapitel 5 – 8) und *Inferenzstatistik* (Kapitel 9 – 12).

Methodenlehre

In Kapitel 1 werden *wissenschaftstheoretische Grundlagen* erläutert. Die Wahl eines geeigneten *Forschungsdesigns*, Kapitel 2, steht am Beginn einer Untersuchung. In Kapitel 3, *Messen*, geht es um die Frage, was unter einer Messung verstanden wird, welche Gütekriterien an eine Messung angelegt werden können und wie man mehrere Messungen zu einem neuen Messinstrument zusammenfassen kann. In den Sozialwissenschaften dominiert nach wie vor die Befragung, der aus diesem Grunde der zentrale Platz in Kapitel 4, *Erhebungsmethoden*, eingeräumt wurde.

Statistik

Die Statistik läßt sich allgemein in einen *deskriptiven* und einen *inferenzstatistischen* Teil gliedern. Mit deskriptiven Statistiken werden vorliegende Daten beschrieben. Die Inferenzstatistik zielt darauf ab, mit Daten einer Auswahl (*Stichprobe*) auf eine größere Gesamtheit zu schließen.

Den *deskriptiven* Teil beginnen wir mit einem Kapitel zu *Tabellen und Graphiken* (Kapitel 5). Mit *Mittel- und Streuungswerten* (Kapitel 6) werden Verteilungen von Merkmalen charakterisiert. Die Stärke der Beziehung zwischen zwei Merkmalen kann mit *Zusammenhangsmaßen* (Kapitel 7) ausgedrückt werden, während die *lineare Einfachregression* (Kapitel 8) es ermöglicht, die Größe des (linearen) Einflusses eines Merkmals auf ein anderes zu berechnen.

Den Auftakt zum *inferenzstatistischen* Teil bildet Kapitel 9, in dem wir Möglichkeiten darstellen, per *Auswahlverfahren* Stichproben zu ziehen, die Aussagen über eine Grundgesamtheit erlauben. Grundlage solcher Schlüsse sind *Wahrscheinlichkeitsverteilungen* (Kapitel 10). Mit Konfidenzintervallen, Kapiteln 11, schätzen wir Parameter der Grundgesamtheit auf Basis einer Stichprobe. *Testverfahren*, Kapitel 12, dienen dazu, Hypothesen über eine Grundgesamtheit an einer einzigen Stichprobe zu testen.

In *Anhang A* finden sich die für die Inferenzstatistik notwendigen z-, t- und χ^2-Tabellen. In *Anhang B* bieten wir die Lösungen zu den Übungsaufgaben, die sich am Ende jedes Kapitels befinden. Diese Aufgaben sind dazu gedacht, sich über die wichtigsten Punkte jedes Kapitels nochmals Klarheit zu verschaffen. Ein *Register* soll helfen, schnelle Antworten auf konkrete Fragen zu bekommen. Schließlich bieten wir mit der *Online-Unterstützung* zahlreiche weitere Informationen an, die von den auf Seite x genannten *WWW*-Servern bezogen werden können.

Das Buch ist so aufgebaut, dass alle Berechnungen von Hand bzw. mit einem Taschenrechner nachvollzogen werden können. Unsere Erfahrung mit Statistikkursen und Einführungen in SPSS und Stata zeigt, dass Probleme weniger in der Bedienung der Software (vgl. zu SPSS: Brosius 2006; zu Stata: Kohler und Kreuter 2008) als vielmehr im Verständnis der statistischen Verfahren bestehen. Für diejenigen, die die Beispiele mit SPSS oder Stata nachrechnen wollen, haben wir die Datensätze auf der Internetseite des Buches zur Verfügung gestellt.

Das Buch wurde mit dem Textsatzsystem LaTeX 2ε gesetzt, die Graphiken wurden weitgehend mit dem Programm TDA von Götz Rohwer und Ulrich Pötter – eigentlich ein Programm zur Ereignisdatenanalyse – erzeugt. Außerdem haben wir die Statistik-Lernprogramme GSTAT und GSTAT2 von Fred Böker verwandt, mit denen die Grundlagen der Inferenzstatistik auf einfache Art und Weise nachvollzogen werden können. Alle genannten Programme sind frei erhältliche Software: LaTeX 2ε ist u. a. auf der Seite http://www.dante.de erhältlich, TDA wird unter http://www.stat.ruhr-uni-bochum.de/tda.html zur Verfügung gestellt. GSTAT und GSTAT2 finden sich im Internet unter http://www.stat-oek.wiso.uni-goettingen.de/user/fred/; die dazugehörige Literatur kann preisgünstig über den Buchhandel bezogen werden (vgl. Böker 1993, 1998).

In der aktuellen Auflage haben wir den Titel geändert. Wir tragen damit der Tatsache Rechnung, dass das Buch gleichermaßen für Lehrveranstaltungen in der Politikwissenschaft *und* der Soziologie verwendet wird. Die inhaltlichen Beispiele stammen vorwiegend aus der politischen Soziologie, wodurch der Anwendungsbezug für Studierende beider Fächer gegeben ist.

Bei den Teilnehmerinnen und Teilnehmern unserer Statistik-Veranstaltungen an den Universitäten Trier, Mainz, Jena und Siegen möchten wir uns für deren konstruktive Hinweise bedanken. Das Buch hat von den detaillierten Anmerkungen durch Hans-Jürgen Andreß zur ersten Auflage wesentlich profitiert. Hilfe bei der Überarbeitung der verschiedenen Auflagen leisteten Matthias Pflume, Ulrich Teusch, Martina Eltges, Andrea Schulze, Thomas Lenz und Nicole Zillien.

Es gibt Menschen, die gerne im Duden schmökern: Matthias P. Heck, Mainz, hat eine Verwendung der neuen deutschen Rechtschreibung schon bei der ersten Auflage angemahnt und diese bei der aktuellen Auflage dann auch tatkräftig unterstützt.

Trier, Mai 2009

Cornelia Weins Uwe Gehring

Online-Unterstützung

Auf den unten genannten *WWW*-Seiten bieten wir Materialien und ergänzende Informationen an. Unter anderem findet sich dort eine Formelsammlung inklusive der Tabellen aus Anhang A. Die Tabellen, Abbildungen und Formeln dieses Buches können in stark vergrößerter Form (zum Beispiel für Folien) kopiert werden. Für diejenigen, die die Beispiele mit Hilfe von SPSS, SAS oder Stata nachrechnen wollen, stehen Datensätze bereit. Schließlich verweisen „Links" auf weitere Informationen im Netz.

`http://www.grundkurs-statistik.de`

Inhalt

Tabellenverzeichnis

Abbildungsverzeichnis

1 Wissenschaftstheoretische Grundlagen

1.1 Wissenschaftstheorien

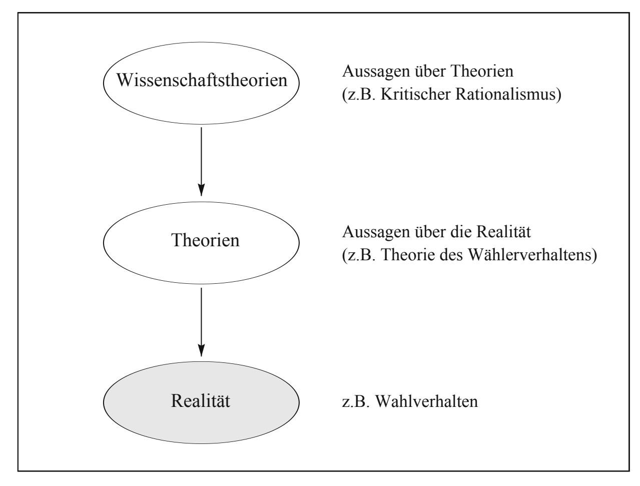
Es mag vielleicht verwundern, dass ein Buch zur empirischen Sozialforschung und Statistik mit einem Kapitel zur Wissenschaftstheorie beginnt. Da die Vorgehensweise einer Untersuchung jedoch vom wissenschaftstheoretischen Blickwinkel geprägt wird, werden wir dieses Kapitel vor allem dazu nutzen, eine sehr bekannte wissenschaftstheoretische Position, den Kritischen Rationalismus, zu skizzieren. Grundlegende Einführungen in die Wissenschaftstheorie bieten Poser (2001), Ritsert (2003) und Chalmers (2007).

Zunächst dazu, was Wissenschaft überhaupt ist. Mit Hilfe von Wissenschaft versuchen Menschen, ihre Erkenntnis über die Realität zu erweitern (vgl. Patzelt 2007, 67). Was passiert bei der Photosynthese, wie entstehen Sterne, was sind die Ursachen gesellschaftlicher Ungleichheit und welche Gründe sind ausschlaggebend für die Wahl einer bestimmten Partei? Wissenschaftliche Erkenntnisgewinnung unterscheidet sich dabei nicht prinzipiell davon, wie man im Alltagsleben Wissen erwirbt. Folgert man aus der Beobachtung, dass das morgendliche Frühstücksei immer dann hart wird, wenn man es zehn Minuten lang kocht, dass alle Eier nach einem zehnminütigen Kochvorgang hart sind, so hat man sein Wissen auf die gleiche Art und Weise (nämlich durch Verallgemeinerung) erweitert wie der Wissenschaftler, der mehrmals nach Zugabe einer Substanz zu einer anderen die gleiche chemische Reaktion beobachtet und daraus ableitet, dass diese Reaktion immer stattfindet. Im Gegensatz zum Alltagswissen zeichnet sich Wissenschaft jedoch durch einen höheren Abstraktionsgrad, ein systematischeres Vorgehen und vor allem die kritische Überprüfung der gewonnenen Erkenntnisse aus. Leider ist es jedoch – wie wir später noch sehen werden – auch mit Hilfe von Wissenschaft nicht möglich zu überprüfen, ob die gewonnenen Erkenntnisse wahr sind.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_1,

Wissenschaftstheorien (der Plural zeigt schon an, dass es mehrere gibt) sind *Aussagenbündel darüber, was Wissenschaft ist* und wie diese vorzugehen hat. Sie sind also noch keine Theorien über einen Ausschnitt der Realität (also z. B. die Sternentstehung oder das Wahlverhalten), sondern *Theorien über Theorien*, die auch als *Metatheorien* bezeichnet werden. Wissenschaftstheorien bestimmen also, wie die eigentlich interessierende Theorie über die Realität auszusehen hat. Sie beschäftigen sich mit Fragen wie: Welche Aussagen sind in Theorien zulässig, welche Methoden werden angewendet, welche Ziele verfolgt Wissenschaft? Sind normative (wertende) Aussagen in der Wissenschaft erlaubt oder nicht? Und: Welche Rolle spielen Werte in der Wissenschaft. Theorien treffen dagegen Aussagen über einen Ausschnitt der Realität, eine Theorie des Wählerverhaltens also über das Wahlverhalten, wie es in Abbildung 1.1 zum Ausdruck kommt.

Abbildung 1.1: Der Status von Theorien



In der deutschen Politikwissenschaft wurde lange Zeit eine Unterscheidung der wissenschaftstheoretischen Positionen in *normativ-ontologisch*, *empirisch-analytisch* und *kritisch-dialektisch* vorgenommen (vgl. zur Einführung Druwe 1994, 57-74). Diese Ansätze unterscheiden sich vor allem in Bezug auf den Erkenntnisgegenstand (Was soll erkannt und erklärt werden?) und die Erkenntnisquelle (Empirie oder Vernunft?). Die *normativ-*

ontologische Wissenschaft versucht, das *Wesen* ihres Gegenstandes, also z. B. des Staates oder der Gesellschaft, zu erfassen. Ziel der Ontologie ist es, das Wesen, d. h. den Idealzustand eines Gegenstandes, zu erkennen und daraus Handlungsanleitungen abzuleiten. Im Vordergrund steht das Nachdenken über Politik mittels philosophischer Reflexionen, wozu vor allem die Hermeneutik und die Phänomenologie dienen. Die *kritisch-dialektische Wissenschaft* stellt dagegen die *Totalität der Gesellschaft* und die *Emanzipation des Menschen* aus Abhängigkeiten in den Vordergrund. Ziel ist es, mittels dialektischer, aber auch hermeneutischer und empirischer Methoden, Herrschaftsverhältnisse offenzulegen und Gesellschaftskritik zu üben. Hierzu zählt die kritische Theorie von Adorno und Habermas. Die *empirisch-analytische Wissenschaft* versucht demgegenüber, aufgrund von beobachtbarer (= empirischer) Realität *Gesetzmäßigkeiten sozialen Handelns* zu erkennen. Ihr Ziel ist die Beschreibung, Erklärung und Prognose sozialer Tatbestände mit den Techniken der empirischen Sozialforschung. Zu diesem Ansatz gehören z. B. der *Positivismus* und der *Kritische Rationalismus* (siehe Kapitel 1.2).

Diese aus der Politikwissenschaft stammende Einteilung wissenschaftstheoretischer Ansätze sollte allerdings lediglich als grobes Raster angesehen werden. Die Klassifikation der Ansätze ist allein deshalb problematisch, weil ihre Einteilung nach Kriterien erfolgt, die auf unterschiedlichen Ebenen angesiedelt sind. Ontologische und kritisch-dialektische Ansätze werden nach dem Ziel von Wissenschaft definiert, empirisch-analytische nach der Methode. Bedeutsame wissenschaftstheoretische Standpunkte, wie etwa der *Konstruktivismus* (vgl. Ritsert 2003, Kapitel 7), werden zudem nicht erfasst.

Die Unterschiede zwischen verschiedenen wissenschaftstheoretischen Standpunkten kristallisierten sich im 20. Jahrhundert in der *Werturteilsdebatte* und im *Positivismusstreit* heraus (vgl. Ritsert 2003, 65-140). Im Werturteilsstreit wurde im Verein für Socialpolitik vor allem um die Frage gestritten, welche Rolle Werte innerhalb der Wissenschaft spielen. Max Weber vertrat in dieser Auseinandersetzung das *Wertfreiheitspostulat*, die Forderung der Trennung von Werturteilen und wissenschaftlichen Aussagen. Im Positivismusstreit wurde die Auseinandersetzung zwischen Karl Popper und Hans Albert als Vertretern des kritischen Rationalismus auf der einen Seite und Theodor Adorno und Jürgen Habermas als Vertretern der kritischen Theorie auf der anderen Seite geführt, wobei die Differenzen nicht einfach bestimmbar sind. Im Kern ging es auch

hier um die Rolle von Werten in der Wissenschaft und die Aufgabe von Wissenschaft. Wissenschaftliche Kritik ist für Vertreter der kritischen Theorie immer auch Gesellschaftskritik.

1.2 Das Forschungsprogramm des Kritischen Rationalismus

Der Kritische Rationalismus ist ein von Karl Raimund Popper begründetes Forschungsprogramm (vgl. Popper 1971). Der zentrale wissenschaftstheoretische Beitrag besteht in der Aufgabe des Rechtfertigungsgedankens von Theorien zugunsten des Falsifikationsprinzips. Mit Rechtfertigungsgedanke ist der Versuch gemeint, Theorien durch ihren Wahrheitsgehalt zu bestätigen. Wahrheit bedeutet dabei nichts anderes als die Übereinstimmung mit der Realität. Wir werden die Popper'sche Wende in der Wissenschaftstheorie ausgehend von der Struktur einer wissenschaftlichen Erklärung erläutern.

1.2.1 Die Struktur einer wissenschaftlichen Erklärung

Ein Ereignis zu erklären heißt, dass wir es aus Gesetzen und Randbedingungen deduktiv ableiten. Das von Popper gewählte Beispiel zur Illustration einer wissenschaftlichen Erklärung ist ein Faden, der reißt. Kausal ist das Reißen des Fadens erklärt, wenn man weiß, dass der Faden eine Reißfestigkeit von 1 kg besitzt, aber mit einem Gewicht von 2 kg belastet wurde. Die Erklärung beinhaltet ein Gesetz („Jedesmal, wenn ein Faden mit einer Last von einer gewissen Mindestgröße belastet wird, zerreißt er") und Randbedingungen („Für diesen Faden hier beträgt diese Größe 1 kg" und „Das an diesen Faden angehängte Gewicht ist ein 2-kg-Gewicht") (Popper 1971, 31 f.).

Weil durch logische Ableitung (deduktiv) von einem Gesetz und den Randbedingungen auf das zu erklärende Ereignis (das Reißen des Fadens) geschlossen wird, heißt dieses Modell *deduktiv-nomologische* Erklärung. Die Struktur dieser Erklärung wurde von Hempel und Oppenheim (1948) herausgearbeitet, weshalb diese auch als *H-O-Schema* bezeichnet wird (vgl. Opp 2005, 46-52).

Tabelle 1.1: Deduktiv-nomologische Erklärung

Gesetz	Randbedingung	Zu erklärende Beobachtung
Alle Rotdrosseln wandern	Vogel X ist eine Rotdrossel	Vogel X wandert
Alle Arbeiter wählen SPD	Person Y ist Arbeiter	Person Y wählt die SPD
Explanans **Prämissen**		Explanandum **Konklusion**

Eine deduktiv-nomologische Erklärung besteht also aus mindestens einem *Gesetz*, mindestens einer *Randbedingung* und einer Beobachtung, die erklärt werden soll (*zu erklärendes Ereignis*). Gesetze sind deterministische, räumlich und zeitlich unbegrenzte Aussagen. Deterministisch bedeutet, dass Gesetze als All-Aussagen formuliert werden können. Allen Elementen der interessierenden Menge wird eine Eigenschaft zugeschrieben. Gesetze und Randbedingungen werden auch als Explanans oder Prämissen bezeichnet, der zu erklärende Sachverhalt als Explanandum oder Konklusion. Anhand der beiden Beispiele in Tabelle 1.1 kann das Prinzip verdeutlicht werden.

Im ersten Beispiel möchten wir erklären, warum ein bestimmter Vogel im Winter die südlichen Gefilde bevorzugt. Beispielsweise könnten wir den Vogel in unserem Garten als Rotdrossel identifizieren und uns noch dunkel aus dem Biologieunterricht daran erinnern, dass alle Rotdrosseln Zugvögel sind. Der beobachtete Vogel fliegt also im Winter in den Süden, weil er eine Rotdrossel ist. Im zweiten Beispiel lautet das Gesetz, dass alle Arbeiter SPD wählen. Mit diesem Gesetz und der Randbedingung, dass Y ein Arbeiter ist, können wir die Wahl der SPD durch Person Y erklären.

Aus wahren Prämissen lassen sich mit Hilfe der deduktiven Logik wahre Schlüsse ableiten. Das Umgekehrte gilt nicht. Aus empirisch zutreffenden Schlussfolgerungen kann nicht die Wahrheit der Prämissen gefolgert werden. Beispielsweise könnte der beobachtete Vogel zwar wandern, aber nicht deshalb, weil er eine Rotdrossel ist, sondern weil er zu einer anderen Sorte Zugvögel zählt. Das herangezogene Gesetz wäre in diesem Fall also nicht die richtige Erklärung für das Zugverhalten des Vogels, weil es sich nicht um eine Rotdrossel handelt.

Die Gültigkeit einer wissenschaftlichen Erklärung beruht demnach auf der *Wahrheit* der im Explanans verwendeten Aussagen. Für die wissenschaftstheoretische Auseinandersetzung spielte vor allem die Forderung wahrer Gesetze eine große Rolle. Eine Möglichkeit, die immer wieder in Betracht gezogen wurde, ist die Verifikation von Gesetzen durch *Induktion* (vgl. Poser 2001, 108-119). Die Induktion ist ebenso wie die Deduktion ein Schlussverfahren. Allerdings werden bei der Induktion aus singulären Aussagen („Dieser Schwan ist weiß“, „der Schwan dort hinten ist weiß“ usw.) allgemeine Aussagen („Alle Schwäne sind weiß“) abgeleitet. Im Gegensatz zu deduktiven Schlüssen sind induktive Schlüsse gehaltserweiternd, weil wir auf eine größere Zahl von Fällen schließen. Die Verifikation von Gesetzen durch Induktion hat allerdings einen Haken, der seit Hume als das *Induktionsproblem* bekannt ist. Auch wenn wir bisher immer nur weiße Schwäne gesehen haben, können wir daraus nicht folgern, dass dies auch in Zukunft so sein wird. Zudem ist es möglich, dass es nicht-weiße Schwäne gibt oder gab, die wir nicht beobachten bzw. beobachtet haben. Die ernüchternde Antwort ist daher: Wir können die Wahrheit von Gesetzen nicht durch Induktion beweisen. Das heißt natürlich nicht, dass ein Gesetz nicht wahr sein kann; die Wahrheit ist aber nicht feststellbar.

Popper hat dies klar erkannt und zeigt uns eine Alternative auf. Ein Gesetz lässt sich zwar niemals durch Beobachtungen bewahrheiten (*verifizieren*). Eine einzige widersprechende Beobachtung reicht jedoch aus, um eine gesetzesartige Aussage zu widerlegen (*falsifizieren*).

1.2.2 Falsifikation statt Induktion

Wegen des Induktionsproblems schlägt Popper vor, alle „Gesetze“ strikt als Hypothesen aufzufassen. Als Möglichkeit der Falsifikation von Hypothesen kann das *H-O-Schema* verwendet werden: Aus der Hypothese und den Randbedingungen werden Beobachtungssätze abgeleitet, die im kritischen Rationalismus *Basissätze* genannt werden. Erweist sich ein Basissatz als falsch, so wird die Hypothese widerlegt. Sobald wir eine einzige Rotdrossel entdecken, die im Winter hier bleibt, *wissen* wir, dass nicht alle Rotdrosseln Zugvögel sind. Vorausgesetzt, es handelt sich tatsächlich um eine Rotdrossel. Allgemein formuliert: Liegt die Randbedingung vor, nicht aber die Schlussfolgerung, dann ist die Hypothese widerlegt.

Wir gehen so lange von der Gültigkeit der Hypothese aus, bis diese sich als falsch erwiesen hat. Solange wir ausschließlich Rotdrosseln beobach-

ten, die im Winter in den Süden fliegen, gehen wir also von der Gültigkeit der Aussage „Alle Rotdrosseln wandern“ aus. Eine Hypothese wird beibehalten, wenn die Randbedingung und der Basissatz zutreffen. Gleichzeitig versuchen wir, unsere Hypothese wiederholt an der Realität zu prüfen. Eine Theorie hat sich nach Popper *bewährt*, wenn sie mehreren strengen Prüfungen standgehalten hat. Hält eine Hypothese einer Prüfung nicht stand, so muss sie verworfen und durch eine neue ersetzt werden.

Falsifikation wird erst möglich, wenn die Hypothesen auch tatsächlich „an der Erfahrung scheitern können“ (vgl. Popper 1971, 15). Die Aussagen, die in Hypothesen und Gesetzen verwendet werden, müssen *empirischen* Gehalt haben. Die Aussage „Wer Böses tut, landet in der Hölle“ beruht wie alle *metaphysischen Aussagen* nicht auf Erfahrung und kann daher auch nicht durch Erfahrung widerlegt werden. An der Realität können ebenso wenig Aussagen scheitern, die immer wahr sind. Solche Aussagen werden als *Tautologien* bezeichnet. Ein Beispiel für eine *Tautologie* wäre: „Nach dem Lesen des Kapitels zur Wissenschaftstheorie verstehen Sie das Falsifikationsprinzip oder Sie verstehen es nicht.“ Dieser Satz ist immer wahr, denn die Folgerung beinhaltet alle möglichen Ereignisse. In diesem trivialen Beispiel ist die Tautologie natürlich leicht zu entdecken. In sozialwissenschaftlichen Theorien kann das schon schwieriger sein (vgl. Diekmann 2008, 157 f.). Die potenzielle Falsifizierbarkeit von Hypothesen grenzt empirische Wissenschaften von nicht-empirischen Wissenschaften ab.

Nicht widerlegbar sind auch *Existenzaussagen*, d. h. Aussagen über das Vorhandensein eines Gegenstandes oder mehrerer Gegenstände. Eine mögliche Existenzaussage lautet etwa: „Es gibt einen weißen Schwan.“ Wollten wir diese Aussage widerlegen, so müssten wir die Farbe aller Schwäne in Vergangenheit, Gegenwart und Zukunft kennen. Umgekehrt reicht bereits ein einziger zutreffender Fall zur Bewahrheitung einer Existenzaussage. Sehen wir einen weißen Schwan, dann ist die Aussage verifiziert. An diesem Beispiel zeigt sich die *Asymmetrie zwischen Falsifikation und Verifikation* bei All-Aussagen und Existenzaussagen. Bei einer Existenzaussage genügt eine zutreffende Beobachtung zur Verifikation, während diese nicht widerlegt werden kann. Genau umgekehrt verhält es sich mit den in Hypothesen verwendeten All-Aussagen: Hier genügt bereits eine widersprechende Beobachtung zur Falsifikation, während All-Aussagen auch durch noch so viele zutreffende Beobachtungen nicht verifiziert werden können. So wurde die für Europäer durch zahlreiche Beobachtungen belegte Aussage „Alle

Schwäne sind weiß" mit der Entdeckung schwarzer Schwäne in Australien um 1700 widerlegt (vgl. Poser 2001, 111).

Das Modell einer wissenschaftlichen Erklärung geht von raum-zeitlich unbeschränkten All-Aussagen aus. Häufig haben wir es jedoch mit raumzeitlich begrenzten Hypothesen bzw. Theorien zu tun. So behauptet die Theorie des Wertewandels von Ronald Inglehart (Inglehart 1977) in nachindustriellen Gesellschaften ein Wandel von materiellen hin zu postmateriellen Werten (vgl. einführend Bürklin und Klein 1998). Räumlich und zeitlich wird die Theorie auf „nach-industrielle Gesellschaften" beschränkt (was darunter verstanden wird, ist nur eine Definitionsfrage). Über andere Gesellschaften wird also zunächst keine Aussage getroffen. Der Informationsgehalt der Theorie wird durch die Eingrenzung geringer. Die Gefahr besteht darin, dass die raum-zeitliche Eingrenzung einer Theorie so weit geht, dass es keine potenziellen Falsifikatoren mehr für die Theorie gibt. Die Theorie wäre dann gegenüber Kritik *immunisiert*.

1.2.3 Basissatzproblem

Die Falsifikation von Hypothesen ist mit einem Problem konfrontiert, das Popper das *Basissatzproblem* genannt hat. Wir haben gesagt, dass eine Hypothese dann falsifiziert ist, wenn sie einer Konfrontation mit der Realität nicht standhält. Eine Hypothese kann jedoch nie direkt durch Beobachtungen geprüft werden, sondern nur anhand einer Aussage über eine Beobachtung. Diese Beobachtungsaussage kann aber fehlerhaft sein. Beobachtungen - selbst Beobachtungen mit bloßem Auge - sind Beobachtungen im Lichte einer Beobachtungs- bzw. Messtheorie. Diese kann sich genauso als falsch erweisen wie die eigentlich interessierende Theorie, z. B. weil unser Instrument etwas anderes misst als das, was es messen sollte. Basissätze sind daher „objektiv kritisierbare Prüfsätze" (Popper 1971, 76), deren Wahrheit ebenso wenig bewiesen werden kann, wie die der Theorie selbst.

Aus diesem Dilemma befreit sich die Wissenschaft, indem sie stillschweigend vereinbart, dass der Forscher bei der Überprüfung einer Theorie den höchstmöglichen methodischen Standard einhält und seine Vorgehensweise nachprüfbar und damit der Kritik zugänglich macht. Ist dies der Fall, dann werden die Basissätze vorläufig anerkannt. Die Akzeptanz der Basissätze ist demnach eine konventionelle Festsetzung. Popper hat für das

Problem der schwankenden empirischen Basis ein anschauliches Bild gefunden. Die Wissenschaft ist für ihn ein Bau, dessen Pfeiler nicht auf Fels gründen, sondern sich „von oben her in den Sumpf senken - aber nicht bis zu einem natürlichen ‘gegebenen’ Grund“; „wenn man hofft, daß sie das Gebäude tragen werden, beschließt man, sich vorläufig mit der Festigkeit der Pfeiler zu begnügen“ (Popper 1971, 76).

1.2.4 Probabilistische Hypothesen

Das Falsifikationsprinzip ist zwar eine feine Sache, funktioniert in der beschriebenen Form jedoch nur bei Hypothesen der Form „Immer wenn X, dann Y“. Solche Hypothesen werden als deterministisch bezeichnet. In den Sozialwissenschaften gibt es aber bisher keine deterministischen Hypothesen. Sozialwissenschaftliche Theorien und Hypothesen werden daher als statistische bzw. probabilistische Aussagen formuliert. Wir behaupten nicht mehr, dass alle Arbeiter die SPD wählen, sondern treffen eine Wahrscheinlichkeitsaussage. Beispielsweise in der Form: „Arbeiter stimmen häufiger für die SPD als für jede einzelne andere Partei.“ Bei Arbeitern müsste die SPD demnach die stärkste Partei sein.

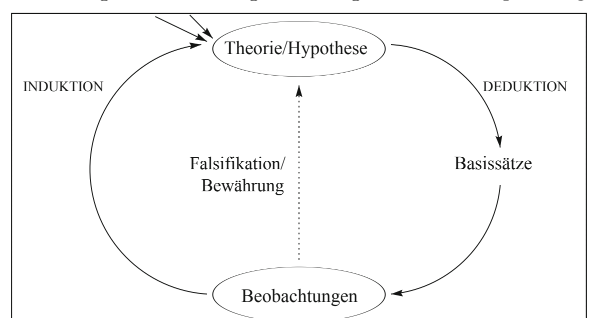
Diese Einschränkung hat erhebliche Konsequenzen. Aus einer probabilistischen Hypothese und den Anfangsbedingungen kann das Explanandum nicht mehr mit Hilfe der deduktiven Logik abgeleitet werden. Ist eine Person Arbeiter, dann ist die Wahrscheinlichkeit einer Wahlentscheidung zugunsten der SPD (bei Gültigkeit der Hypothese) zwar hoch, beträgt aber nicht 100%. Weil das Explanandum nur wahrscheinlich (nicht aber logisch zwingend) ist, sprechen wir hier von einer induktiv-statistischen Erklärung (vgl. Opp 2005, 56 f.). Ein einzelner Arbeiter, der eine andere Partei als die SPD wählt, widerlegt nicht die Hypothese. Probabilistische Hypothesen können daher nicht durch einen einzigen Fall falsifiziert werden. Wir sehen die Hypothese jedoch dann als „falsifiziert“ an, wenn wir bei einer hinreichend großen Zahl von Arbeitern beispielsweise einen höheren Prozentsatz an CDU- als an SPD-Wählern feststellen würden. Diese Akzeptanz einer „Falsifikation“ probabilistischer Hypothesen beruht jedoch auf der Annahme, dass die probabilistische Hypothese für jede beliebige Teilmenge von Fällen gilt (vgl. dazu Prim und Tilmann 1997, 89 ff.).

1.3 Der Ablauf des Forschungsprozesses

1.3.1 „Der Kreis der Wissenschaft“

Den Ablauf des Forschungsprozesses kann man sich wie in der folgenden Abbildung als Kreislauf vorstellen (vgl. Wallace 1971). Wissenschaft ist demnach nichts anderes als eine Verzahnung von *Theoriegewinnung* und *Theorieprüfung* mittels Induktion und Deduktion.

Abbildung 1.2: Theoriegewinnung und Theorieprüfung



Der eigentlich kreative Teil besteht in der Entwicklung von Theorien bzw. Hypothesen. Wie die verschiedenen Pfeile in der Graphik andeuten sollen, kann man auf die unterschiedlichsten Arten zu Theorien gelangen (z. B. durch Nachdenken). Eine gebräuchliche Methode besteht in der Verallgemeinerung einzelner Beobachtungen durch *Induktion*. Wir könnten z. B. eine Vielzahl von Schwänen beobachten. Daraus, dass alle von uns beobachteten Schwäne weiß sind, gelangen wir zu dem Schluss, „Alle Schwäne sind weiß“.

Wir können noch so viele weiße Schwäne beobachten - ein Beweis für die Wahrheit der Theorie ist es nicht, wie wir gesehen haben. Aus diesem Grunde scheidet das Induktionsprinzip zur *Prüfung* einer Theorie aus. Als Möglichkeit der Kritik von Theorien wurde von Popper deshalb das Falsifikationsprinzip vorgeschlagen. Wir leiten aus unserer Theorie Basissätze ab und schauen, ob diese mit unseren Beobachtungen in Einklang stehen oder nicht. Im ersteren Fall hat sich unsere Theorie bewährt, im

letzteren ist sie falsifiziert (gestrichelte Linie). Damit beginnt der Kreislauf von Neuem. Die vorhandenen Beobachtungen können als Grundlage der Modifikation der Theorie oder einer neuen Theorie dienen.

Lakatos (1974) hat darauf hingewiesen, dass Theorien in der Praxis nicht allein aufgrund falsifizierender Beobachtungen aufgegeben werden (können). Dies wäre angesichts des Basissatzproblems und der damit verbundenen Unsicherheit über die empirische Basis auch eine selbstmörderische Strategie. Eine Theorie wird erst dann preisgegeben, wenn wir eine alternative Theorie besitzen, die einen theoretischen Gehaltsüberschuss aufweist (die einen größeren Geltungsbereich als die alte Theorie hat), der zumindest teilweise empirisch bewährt ist (*progressive Problemverschiebung*). Wird eine falsifizierte Theorie mangels geeigneterer Theorien beibehalten bzw. so modifiziert, dass widersprechende Beobachtungen aus deren Geltungsbereich ausgeklammert werden, dann handelt es sich um eine *degenerative Problemverschiebung* (vgl. zur Einführung in die wissenschaftstheoretischen Positionen von Kuhn, Lakatos und Feyerabend Poser 2001).

1.3.2 Der Ablauf einer empirischen Untersuchung

Ausgangspunkt einer Untersuchung ist im Idealfall eine *Theorie*, d. h. ein „*System logisch widerspruchsfreier Aussagen* (Sätze, Hypothesen) über den jeweiligen Untersuchungsgegenstand *mit den zugehörigen Definitionen* der verwendeten Begriffe“ (Kromrey 2006, 52). Logisch konsistent bedeutet, dass sich die Aussagen, die in einer Theorie enthalten sind, nicht widersprechen dürfen. Da es in den Sozialwissenschaften bisher kaum Theorien gibt, liegen den meisten Untersuchungen jedoch *Hypothesen* zugrunde. Eine zu überprüfende Hypothese könnte beispielsweise beinhalten, dass Vorurteile gegenüber Minderheiten mit der Größe der Minderheit in einem Gebiet zunehmen (vgl. bereits Blalock 1967).

Präzisierung der Begriffe

Zur Überprüfung einer Hypothese müssen die verwendeten *Begriffe* zunächst präzise definiert werden. Begriffe sind Mittel, mit deren Hilfe wir das Chaos von Eindrücken sprachlich ordnen. Es sind Regeln zur Strukturierung von Wahrnehmungen. Sie sind dabei niemals identisch mit der Realität, sondern bezeichnen unser Modell von der Realität. Die Bildung

von Begriffen erfordert daher immer ein gewisses Maß an Abstraktion, d. h. eine Theorie darüber, was die gemeinsamen Merkmale des mit dem Begriff Bezeichneten sein sollen.

Begriffe haben unterschiedliche Funktionen: Sie ermöglichen die Kommunikation über Gegenstände und dienen der Klassifikation (vgl. ausführlicher Mayntz et al. 1978, 9–22). Diese Aufgaben von Begriffen sind nicht trivial: Ohne einen Begriff „Vorurteile" ist eine Verständigung über dieses Phänomen nicht möglich. Begriffe, die mehrdeutig oder unbestimmt sind, erfüllen ihre Funktion nicht mehr. Mehrdeutig ist ein Begriff, wenn verschiedene Inhalte mit demselben Begriff bezeichnet werden, z. B. kann mit „Hahn" das Tier oder ein Wasserhahn gemeint sein. Unbestimmt ist ein Begriff, wenn er nicht präzise genug gefasst ist.

Aus diesem Grunde wird die Verwendung von Begriffen in wissenschaftlichen Texten mit Hilfe von *Definitionen* (vgl. Opp 2005, 106-131) festgelegt. *Realdefinitionen* treffen Aussagen über die *Beschaffenheit der Realität*. Versucht wird nach dieser Auffassung, das Wesen eines Gegenstandes zu erfassen. Die Form der Definition entspricht normalerweise einer *Ist*-Aussage. Realdefinitionen beinhalten demnach empirische Aussagen und können daher wahr oder falsch sein. Ein Beispiel wäre die Aussage „Soziologie *ist* eine empirische Wissenschaft". Der Wahrheitsgehalt der Aussage kann beurteilt werden. Würde Soziologie als Wissenschaft sich nicht auf Erfahrung gründen, dann wäre diese Definition falsch. Opp (2005, 113 f.) verweist auf die Mehrdeutigkeit von Wesensbestimmungen (Bedeutungsanalysen, empirische Gesetze, Begriffsexplikationen, normative Aussagen usw.) und rät deshalb von ihrer Verwendung ab.

In den Sozialwissenschaften werden in der Regel Nominaldefinitionen verwendet. Eine *Nominaldefinition* ist eine *Festsetzung der Verwendung eines Begriffes*. Ein Beispiel für eine nominale Definition von Herrschaft findet sich bei Max Weber: „Herrschaft *soll heißen* die Chance, für einen Befehl bestimmten Inhalts bei angebbaren Personen Gehorsam zu finden" (Hervorhebung ergänzt, Weber 1980, 28). Der Begriff dessen Bedeutung festgelegt wird, hier Herrschaft, wird auch als Definiendum bezeichnet, der definierende Ausdruck „Chance ... Gehorsam zu finden" als Definiens. Die Bestandteile des Definiens müssen bekannt sein. Es handelt sich um eine tautologische Umformung, weil das Definiendum dem Definiens gleichgesetzt wird. Statt der Verwendung des Begriffs Herrschaft könnte Max Weber auch immer „Chance ... Gehorsam zu finden" schreiben (was allerdings

etwas umständlich wäre). Eine Nominaldefinition ist eine Festsetzung über die Verwendung eines Begriffs. Sie beinhaltet keine empirische Behauptung und kann daher auch nicht wahr oder falsch sein. Für ein gegebenes Untersuchungsinteresse kann sich eine Nominaldefinition lediglich als mehr oder weniger zweckmäßig erweisen. Max Webers Herrschaftsbegriff bezieht sich ausschließlich auf Interaktionen zwischen Menschen. Herrschaft über Tiere, die Natur etc. wird durch seinen Herrschaftsbegriff nicht erfasst. Eine nominale Definition von Vorurteilen könnte etwa lauten: „Unter Vorurteilen sollen negative Einstellungen gegenüber den Angehörigen einer sozialen Großgruppe verstanden werden“. Vorurteile werden damit als Einstellungen definiert. Diese Definition von Vorurteilen umfasst zudem nur negative Einstellungen.

Operationalisierung

Als Operationalisierung bezeichnet man alle Forschungsoperationen die notwendig sind, um einen Begriff zu messen. Eine Operationalisierung von Vorurteilen gegenüber Ausländern könnte etwa sein: „Je stärker ein Befragter der Aussage 'Wenn Arbeitsplätze knapp werden, dann sollte man die in Deutschland lebenden Ausländer wieder in ihre Heimat schicken' auf einer siebenstufigen Skala zustimmt, umso größer sind dessen Vorurteile gegenüber Ausländern ausgeprägt“. Diese Operationalisierung zielt auf eine Befragung von Personen ab. Der Ausländeranteil kann operationalisiert werden als das Verhältnis der bei den Einwohnemeldeämtern registrierten Personen mit ausländischer Staatsbürgerschaft zu allen bei den Einwohnemeldeämtern registrierten Personen. Hier wird auf amtliche Daten zur Messung zurückgegriffen.

Es gibt für jeden Begriff verschiedene Möglichkeiten der Operationalisierung. Im amerikanischen Kontext wird Bildung häufig als die Zahl der Schuljahre operationalisiert, die eine Person absolviert hat. Eine alternative Operationalisierung ist die CASMIN-Skala (z. B. Lechert et al. 2006). CASMIN steht für Comparative Analysis of Social Mobility in Industrial Nations. CASMIN kombiniert allgemeine und berufliche Zertifikate zu verschiedenen Stufen (Tabelle 1.2).[1] Für die Operationalisierung der Bildung

1 Die Kategorien 2a - 2b - 2c sind vertauscht, weil im deutschen Bildungssystem die allgemeine Sekundarausbildung (Mittlere Reife, 2b) im Anschluss an die allgemeine Schulbildung einer beruflichen Ausbildung (2a) vorausgeht (vgl. Lechert et al. 2006, 4).

Tabelle 1.2: Die CASMIN-Klassifikation

1a	Kein Abschluss
1b	Hauptschulabschluss ohne beruflichen Abschluss
1c	Hauptschulabschluss und berufliche Ausbildung
2b	Mittlere Reife ohne berufliche Ausbildung
2a	Mittlere Reife und berufliche Ausbildung
2c_gen	Fachhochschulreife/Abitur ohne berufliche Ausbildung
2c_voc	Fachhochschulreife/Abitur und berufliche Ausbildung
3a	Fachhochschulabschluss
3b	Universitärer Abschluss

durch die CASMIN-Skala spricht, dass diese die starke Stratifizierung allgemeiner Abschlüsse (Hauptschule, Realschule und Gymnasium) und die berufliche Spezifität des deutschen Bildungssystems angemessen erfasst, die sich unter anderem für die berufliche Erstplatzierung als bedeutsam erwiesen haben (vgl. z. B. Müller und Shavit 1998).

Eine gegebene Operationalisierung kann sich im Verlauf des Forschungsprozesses als angemessen oder unangemessen erweisen. Operationalisierungen müssen daher genauso zur Disposition stehen wie Hypothesen auch. Ist die Datenerhebung abgeschlossen, dann kann der Fehler kaum mehr korrigiert werden.

Erhebung und Auswertung von Daten

Der letzte und für die Prüfung der Theorie entscheidende Schritt besteht in der Erhebung von Daten durch Beobachtung im weiteren Sinne. Die klassischen Formen der Datenerhebung sind die Befragung, die Inhaltsanalyse und die Beobachtung. Man muss jedoch nicht zwangsläufig eigene Daten erheben. Für viele Untersuchungszwecke kann man auf Daten zurückgreifen, die Andere erhoben haben und für wissenschaftliche Analysen zur Verfügung stellen. Wir werden im folgenden Kapitel darauf zurückkommen. Die Daten müssen dann im Hinblick auf die interessierende Hypothese statistisch ausgewertet werden. Voraussetzung für die Akzeptanz der empirischen Resultate ist die Einhaltung der methodischen Standards.

Die Erhebung und die statische Auswertung von Daten sind Gegenstand der weiteren Kapitel dieses Buches.

2 Forschungsdesigns

Bei der Wahl des Forschungsdesigns müssen verschiedene Aspekte beachtet werden: Sollen die Daten selbst erhoben werden oder kann auf bereits vorhandene Daten zurückgegriffen werden? Auf welche Untersuchungsebene zielt die Fragestellung? Welchen Zeitraum sollen die Daten abdecken? Wird ein Experiment oder z. B. eine Umfrage durchgeführt? Da das Forschungsdesign im Nachhinein nicht mehr veränderbar ist, ist es wichtig, eine Untersuchung genau zu planen, um nicht später auf unbrauchbaren „Datenfriedhöfen“ zu sitzen oder Einschränkungen hinsichtlich der Gültigkeit der erzielten Ergebnisse hinnehmen zu müssen.

Daten können als „beobachtete Merkmalsausprägungen auf Merkmalsdimensionen von Untersuchungseinheiten“ (Mayntz et al. 1978, 35) gekennzeichnet werden. Als *Untersuchungseinheiten* (auch: Merkmalsträger) werden die Einheiten bezeichnet, an denen die Beobachtungen vorgenommen werden. Untersuchungseinheiten sind häufig Personen, es kann sich aber auch um Haushalte, Staaten oder andere Einheiten handeln. Merkmalsdimensionen - wir werden im Weiteren den Begriff *Merkmale* verwenden - wären bei Personen z. B. das Geschlecht, das Alter oder das politische Interesse. Die möglichen Kategorien der Merkmale werden als *Merkmalsausprägungen* bezeichnet. Das Merkmal „politisches Interesse“ könnte die Ausprägungen „stark“, „mittel“ und „schwach“ annehmen. *Variablen* sind Merkmale von Untersuchungseinheiten, die mindestens zwei Ausprägungen annehmen können. Sind den Ausprägungen bereits Zahlen zugeordnet worden, etwa 1 für „starkes“, 2 für „mittleres“ und 3 für „schwaches“ politisches Interesse (siehe Kapitel 3), dann werden diese auch *Werte* genannt.

2.1 Datenerhebung

Eine der wichtigsten Entscheidungen, die bei der Planung einer Untersuchung getroffen werden muss, betrifft die Frage, *wer* die zu analysierenden Daten erhebt. Man unterscheidet zwischen:

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_2,

- Primäranalyse
- Sekundäranalyse

Im ersten Fall werden die benötigten Daten selbst erhoben und von demjenigen, der die Daten erhoben hat, auch zuerst ausgewertet – deshalb *Primäranalyse*. Im zweiten Fall wertet man von anderen erhobene und in der Regel bereits ausgewertete Daten ein weiteres Mal aus – deshalb *Sekundäranalyse*. Bei der Sekundäranalyse können die Daten ursprünglich zu einem völlig anderen Zweck erhoben worden sein. Wichtig ist nur, dass sie dem Untersuchungszweck der erneuten Analyse dienlich sind.

Eine Primärerhebung bietet den entscheidenden Vorteil, dass genau die Merkmalsausprägungen erhoben werden können, die benötigt werden. Der Nachteil besteht darin, dass dies mit hohen (manchmal zu hohen) Kosten verbunden sein kann. Bei Sekundäranalysen stehen möglicherweise nicht alle gewünschten Informationen zur Verfügung – dafür hat man nur sehr geringe (oft sogar gar keine) Kosten zu tragen.

Besonders groß ist der Preisunterschied zwischen Primär- und Sekundäranalysen bei Umfragen. Allein die in einer mündlichen Umfrage anfallenden Kosten zur Bezahlung der Interviewer bzw. Portokosten zur Versendung der Fragebögen können erheblich sein. So hat 1996 eine ca. einstündige bevölkerungsweite Befragung mit dem vom Zentrum für Umfragen, Methoden und Analysen (ZUMA)[1] in Mannheim und der Gesellschaft für Marketing-, Kommunikations- und Sozialforschung mbH (GFM-GETAS, heute: IPSOS) in Hamburg gemeinsam durchgeführten SOZIALWISSENSCHAFTENBUS ca. 400.000,- DM gekostet.[2] Die Kosten einer Sekundäranalyse eines vergleichbaren Datensatzes, nämlich des von uns auch in diesem Lehrbuch immer wieder verwendeten ALLBUS 1994, beliefen sich dagegen nur auf 175,- DM (inklusive Codebuch). Für Studierende belaufen sich die Kosten für eine ALLBUS-Umfrage aktuell auf maximal 25 Euro (Daten auf CD).

In Deutschland werden Sekundärdaten von der Abteilung Datenarchiv und Datenanalyse der GESIS (früher: Zentralarchiv für empirische Sozialforschung, ZA) in Köln archiviert und gegen Entgelt für Sekundäranalysen bereitgestellt. Der *Datenbestandskatalog* listet mehrere tausend

1 Heute: Center for Survey Design and Methodology (CSDM) der GESIS.

2 Der SozialwissenschaftenBus wurde zwischen 1985 und 1998 einmal jährlich durchgeführt.

für die wissenschaftliche Öffentlichkeit verfügbare Studien auf. Eine Recherche im Datenbestandskatalog ist über die Internetseite der GESIS (`http://www.gesis.org/dienstleistungen/daten/`) möglich. Über die GESIS können für wissenschaftliche Zwecke unter anderem die Daten des ALLBUS, des Politbarometer und Studien zu Bundes- und Landtagswahlen bezogen werden. Das Angebot ist nicht auf deutsche Studien beschränkt. Es umfasst auch internationale Erhebungen wie den ISSP (International Social Survey Programme), die European Values Study oder das Eurobarometer.

Sekundärdaten sind auch bei anderen Institutionen erhältlich. Die Statistischen Landesämter und das Statistische Bundesamt in Wiesbaden sind eine wichtige Quelle für Sekundärdaten. Dort kann man u. a. Wahl- und Volkszählungsdaten auf unterschiedlichen regionalen Ebenen (z. B. für Verwaltungseinheiten wie Gemeinden und Kreise, aber auch für Landtagswahlkreise usw.) in maschinenlesbarer Form erhalten. Vom Deutschen Institut für Wirtschaftsforschung (DIW) in Berlin wird das Sozio-ökonomische Panel (SOEP) aufbereitet und ebenfalls für wissenschaftliche Zwecke zur Verfügung gestellt. Das Sozio-ökonomische Panel beinhaltet unter anderem detaillierte Indikatoren zur beruflichen und Einkommenssituation der Befragten und ist daher eine wichtige Datenquelle zur Analyse sozialer Ungleichheit. Ein wichtiger Datengeber ist auch das Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit im Institut für Arbeitsmarkt- und Berufssforschung (IAB) in Nürnberg. Das IAB stellt Daten zur Verfügung, die aus den Meldungen der Arbeitgeber an die Sozialversicherungsträger stammen. Solche für Verwaltungszwecke erhobenen Daten nennt man auch *prozessproduzierte Daten*. Daneben erhebt das IAB auch selbst Daten, wie das Betriebspanel. Die Daten des IAB ermöglichen detaillierte Analysen der Erwerbsbiographie von Personen.

2.2 Ebene der Untersuchungseinheit

Nach der *Ebene* auf der die Daten erhoben werden, wird häufig unterschieden zwischen:

- Individualdaten
- Aggregatdaten

Individualdaten beinhalten Informationen über *Personen* bzw. individuelle Merkmale. Beispiele wären die Wahlabsicht oder das Alter von Personen. **Aggregatdaten** beinhalten Informationen über Gruppen bzw. Kollektive, die aus der Zusammenfassung (Aggregation) von Messwerten der Mitglieder dieser Kollektive beruhen, z. B. durch Summenbildung, Prozentuierung oder eine andere Rechenoperation. Bei Aggregatdaten handelt es sich immer um „abgeleitete Daten" (Pappi 1977, 81). Aggregatdaten werden also nicht an den Kollektiven selbst gewonnen, sondern an den Mitgliedern dieser Kollektive. Die Mitglieder der Kollektive können, müssen aber keine Individuen sein. Die Stimmenanteile von Parteien oder das Durchschnittsalter der Bundesdeutschen sind Aggregatdaten, die aus den Individualdaten Wahlabsicht und Alter gebildet wurden. Daten, die *räumlich* aggregiert wurden (z. B. auf der Ebene von Gemeinden, Wahlkreisen oder Staaten), nennt man auch *ökologische Daten*. Individualdaten stammen zumeist aus Umfragen, Aggregatdaten werden häufig von der amtlichen Statistik bereitgestellt.

Gelegentlich werden auch solche Daten von Kollektiven als Aggregatdaten bezeichnet, die nicht auf Aggregation beruhen, wie z. B. ein Index zur Messung von Bürgerrechten in Staaten (vgl. Widmaier 1997, 104). Nach der Klassifikation von Lazarsfeld und Menzel (1972, 228 f.) handelt es sich hierbei um ein globales Merkmal (*global property*) des Kollektivs. Ein anderes Beispiel für ein globales Merkmal wäre das Regierungssystem eines Staates. Globale Merkmale werden direkt am Kollektiv gemessen und besitzen nach der Vorstellung von Lazarsfeld und Menzel (1972) keine Entsprechung auf der Ebene der Mitglieder des Kollektivs. Aggregierte und globale Merkmale sind eine Untergruppe der Kollektivmerkmale (vgl. dazu Pappi 1977, S. 80 f.).

Ob Individualdaten oder Daten für Kollektive erhoben werden, bestimmt sich aus der Forschungsfrage. Geht es um Motive der Wahlentscheidung, so werden sich die Hypothesen möglicherweise auf Eigenschaften einzelner Wähler wie deren Kanzlerpräferenz oder Parteiidentifikation richten. Wenn alle interessierenden Merkmale auf der Individualebene angesiedelt sind, handelt es sich um eine Individualhypothese. Anders verhält es sich dagegen, wenn eine Hypothese über den Einfluss des Katholikenanteils auf den Stimmenanteil der Christdemokraten in Wahlkreisen formuliert wird (vgl. Kapitel 8). Hier bezieht sich die Hypothese auf Kollektive. Dies ist insbesondere in der Vergleichenden Politikwissenschaft und der Internationalen Politik (vgl. Widmaier 1997) häufig der Fall. So beinhaltet beispiels-

weise das viel diskutierte Theorem vom „Demokratischen Frieden“, dass demokratische Staaten keine Kriege gegeneinander führen (vgl. Teusch und Kahl 2001; Chan 1997). Natürlich können sich Hypothesen auch auf Kollektive und Individuen beziehen. Etwa dann, wenn die These beinhaltet, dass die Stimmabgabe zugunsten der CDU bei Katholiken mit zunehmendem Katholikenanteil in der Gemeinde steigt. Oder wenn behauptet wird, dass Vorurteile nicht nur von Individualmerkmalen wie Bildung usw., sondern auch der Größe der Minderheit in einem Gebiet abhängen. Es handelt sich hier um Mehrebenenhypothesen, weil die Kollektivmerkmale (Katholikenanteil, Größe der Minderheit) und individuelle Merkmale (Konfession, Bildung) als relevant zur Erklärung des Verhaltens (Stimmabgabe) bzw. der Einstellungen (Vorurteile) angesehen werden. Mehrebenenanalysen (vgl. Ditton 1998; Snijders und Bosker 1999) erlauben die simultane Analyse von Daten auf verschiedenen Ebenen. Mehrebenenanalysen setzen voraus, dass hinreichend viele Individuen *und* Kollektive untersucht werden. Nach Snijders und Bosker (1999, 140, 150) sollten auf jeder Ebene mindestens 30 Einheiten untersucht werden. Bei Analysen in denen Staaten die Kollektive sind, ist diese Bedingung für die Ebene der Staaten häufig nicht erfüllt. Zudem muss sichergestellt sein, dass auch die Kollektivmerkmale genügend Varianz aufweisen. So sollte sich beispielsweise der Katholikenanteil in den untersuchten Kollektiven unterscheiden.

Aggregatdaten können auf einem unterschiedlichen *Aggregationsniveau* vorliegen; Bundestagswahlergebnisse beispielsweise auf Ebene der Bundestagswahlkreise, auf Ebene der Bundesländer oder auf Bundesebene. Von den Statistischen Ämtern werden z. B. die ursprünglich als Individualdaten vorliegenden Volkszählungsdaten (Geschlecht, Religionszugehörigkeit, Berufszugehörigkeit, Schulabschluss usw.) auf verschiedenen Ebenen (Gemeinden, Kreise usw.) aggregiert und auch nur in aggregierter Form weitergegeben. Aus naheliegenden Gründen werden auch Wahldaten nur als Aggregatdaten zur Verfügung gestellt. Mit der Aggregation ist ein Informationsverlust verbunden. Bezogen auf das Volkszählungsbeispiel: die aggregierten Volkszählungsdaten geben lediglich Auskunft über die Anzahl der Männer, der Frauen, der Menschen mit einem bestimmten Schulabschluss usw. in einem bestimmten Gebiet. Wie viele Frauen und wie viele Männer welchen Schulabschluss haben, lässt sich den aggregierten Volkszählungsdaten nicht mehr entnehmen. Aus den aggregierten Volkszählungsdaten lassen sich die ursprünglichen Individualdaten nicht mehr herstellen, eine Disaggregation der Daten ist nicht möglich.

Wie das Beispiel der Volkszählungsdaten zeigt, welche nur als Aggregatdaten zugänglich gemacht werden, kann man sich nicht immer aussuchen, ob man mit Individual- oder Aggregatdaten arbeitet. So ist die historische Wahlforschung weitgehend auf die Analyse von Aggregatdaten angewiesen, da bis in die 50er Jahre des 20. Jahrhunderts Umfragedaten sehr rar sind. Für Analysen der Wahlen des Deutschen Reichs greift man auf die Volkszählungs- und Wahldaten zurück, die in der *Statistik des Deutschen Reichs* veröffentlicht wurden (vgl. exemplarisch Winkler 1995; Falter 1991). Es kann also passieren, dass man Aussagen über Individuen treffen möchte, tatsächlich aber nur Aggregatdaten zur Verfügung stehen.

Ökologischer Fehlschluss

In diesem Zusammenhang muss man darauf achten, keinen *Fehlschluss* zu begehen. *Fehlschlüsse entstehen immer dann, wenn Aussageeinheit und Untersuchungseinheit auf unterschiedlichen Ebenen angesiedelt sind.* Schließt man von Beziehungen auf der Aggregatebene auf Beziehungen der Individualebene (allgemein: einer niedrigeren Ebene), begeht man einen *ökologischen Fehlschluss*. Schließt man im umgekehrten Fall von Beziehungen auf der Individualebene auf Beziehungen der Aggregatebene, liegt ein *individualistischer Fehlschluss* vor.

Für die Sozialwissenschaften ist vor allem der **ökologische Fehlschluss** (vgl. Robinson 1950) von Bedeutung, da die Daten häufig in stärker aggregierter Form vorliegen, als man sie für die beabsichtigten Aussagen bräuchte. So schlossen einige Historiker (vgl. die Literaturhinweise bei Falter et al. 1983, 528) aus dem bei den Reichstagswahlen zwischen 1930 und 1932 zeitgleich erfolgten Anstieg der Arbeitslosigkeit und den Wahlerfolgen der NSDAP auf *Reichsebene*, dass Arbeitslose überproportional häufig für die NSDAP gestimmt hätten.

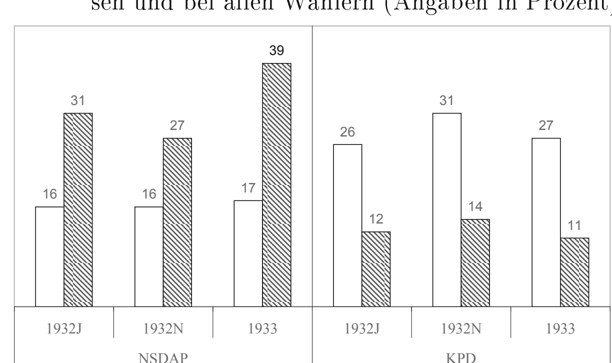
Aufgrund von Zusammenhängen auf der Reichsebene (Zahl der Arbeitslosen und NSDAP-Stimmen) wurden Aussagen über Zusammenhänge auf der individuellen Ebene (Arbeitslosig*keit*und NSDAP-Stimm*abgabe*) getroffen. Dieser Schluss ist jedoch nicht zulässig und kann sich auch inhaltlich als falsch erweisen, nämlich dann, wenn Arbeitslose nicht überproportional häufig für die NSDAP gestimmt haben (vgl. auch Frey und Weck 1981, 6 und 25). Auf der Individualebene lässt sich der Zusammenhang nicht mehr untersuchen. Allerdings kann man den Zusammenhang auf einem niedrigeren Aggregationsniveau prüfen. So stellten Falter et al. (1983)

fest, dass die NSDAP in *Kreisen* mit einem hohen Anteil erwerbsloser Angestellter und Arbeiter durchschnittlich keine höheren Stimmenanteile erzielte. Im Gegenteil: In Kreisen mit einem hohen Erwerbslosenanteil schnitt die NSDAP durchschnittlich sogar schlechter ab (vgl. Falter et al. 1983, 532). Die Ergebnisse von Falter et al. sprechen gegen den auf Reichs- und Bezirksebene (vgl. Frey und Weck 1981) festgestellten positiven Einfluss der Erwerbslosigkeit auf den Stimmenzuwachs der NSDAP zwischen 1930 und 1932, da die Erklärungskraft auf Kreisebene höher ist als auf der Ebene der Bezirke bzw. des gesamten Reiches. Dennoch kann man auch von den weniger stark aggregierten Kreisdaten nicht einfach auf das Wahlverhalten von Arbeitslosen schließen.

Zwar ist das Problem des ökologischen Fehlschlusses nicht lösbar; es sind jedoch statistische Verfahren entwickelt worden (*ökologische Inferenz*), mit deren Hilfe Zusammenhänge der individuellen Ebene *geschätzt* werden können. Mit einer *ökologischen Regression* zeigt Falter (1991), dass die Stimmenanteile für die NSDAP bei Arbeitslosen unterdurchschnittlich waren. Arbeitslose scheinen in deutlich stärkerem Umfang für die KPD als die NSDAP gestimmt zu haben (vgl. Abbildung 2.1). Zumindest direkt scheint die NSDAP nicht von der Massenarbeitslosigkeit profitiert zu haben. Die neueren Verfahren (vgl. King 1997) kombinieren die von Falter verwendete Methode der ökologischen Regression mit der Methode der Grenzen (*method of bounds*) und treffen weniger restriktive Annahmen zur Schätzung der Anteile (vgl. einführend Gschwend 2005). Ohne Hinzuziehung solcher Verfahren sollten die Aussagen einer Untersuchung sich immer auf die Ebene der Untersuchungseinheit (Analyseebene) beziehen, nie auf eine andere Ebene.

An einem weiteren gern verwendeten Beispiel (vgl. Bürklin und Klein 1998, 35 f.) lässt sich die Problematik des ökologischen Fehlschlusses besonders gut verdeutlichen: Bei Bundestagswahlen besteht auf Wahlkreisebene ein positiver Zusammenhang zwischen dem Anteil der Ausländer und dem Stimmenanteil der GRÜNEN. Dennoch würde niemand aus diesem Ergebnis folgern, dass Ausländer in hohem Umfang die GRÜNEN wählen, da Ausländer auf Bundesebene kein Wahlrecht besitzen. Die Aussage „Je höher der Ausländeranteil in einem Wahlkreis, umso besser schneiden die GRÜNEN ab“ ist richtig; die Aussage „Ausländer wählen überproportional häufig die GRÜNEN“ dagegen offenkundig falsch.

Abbildung 2.1: Stimmenanteile der NSDAP und der KPD bei Arbeitslosen und bei allen Wählern (Angaben in Prozent)



Quelle: Falter 1991, 311. 1932J: Juli 1932, 1932N: November 1932

2.3 Untersuchungsanordnung

Generell kann man zwischen zwei Gruppen von Untersuchungsanordnungen unterscheiden:

- Experimente
- Ex-post-facto-Anordnungen

In **Experimenten** werden die Bedingungen der Untersuchung selbst hergestellt und unterliegen damit der Kontrolle und der Einflussnahme des Forschers (vgl. Sarris 1999). Experimentelle Anordnungen sehen bei zwei Gruppen so aus, dass einer Gruppe eine Behandlung zuteil wird (*Experimentalgruppe*), einer zweiten Gruppe - die sich ansonsten von der ersten Gruppe nicht unterscheidet - jedoch nicht (*Kontrollgruppe*). Wie die Behandlung (*treatment* oder *Stimulus*) wirkt, kann nach dem Versuch an möglichen Unterschieden zwischen Experimental- und Kontrollgruppe abgelesen werden. Der Stimulus stellt die unabhängige Variable dar; die Variable also, von der ein Einfluss auf das interessierende Merkmal (die abhängige Variable) vermutet wird.

Um die Unterschiede zwischen Experimental- und Kontrollgruppe auszuschalten, werden die zur Verfügung stehenden Probanden nach einem Zufallsverfahren den Gruppen zugewiesen; dieses Verfahren nennt man *Randomisierung* oder Zufallsaufteilung. Randomisierte Gruppen dürften sich in der Verteilung ihrer Eigenschaften, einmal abgesehen von Zufallsschwankungen, nicht unterscheiden. Durch die Randomisierung wird sichergestellt, dass mögliche Unterschiede zwischen Kontroll- und Experimentalgruppe *nach* dem Experiment tatsächlich auf den Stimulus zurückzuführen sind und nicht etwa aus der unterschiedlichen Zusammensetzung der beiden Gruppen resultieren. Die Ursache-Wirkungs-Beziehung (*Kausalität* ist damit feststellbar. Zusätzlich zur Beobachtung nach dem Versuch kann eine Beobachtung vor dem Versuch stattfinden (*Vorher-Nachher-Messung*). Dadurch kann kontrolliert werden, ob sich die beiden Gruppen in dem interessierenden Merkmal bereits vor dem Setzen des Stimulus (T) unterscheiden. Durch die Vorher-Messung können jedoch Lerneffekte bei den Teilnehmern der Untersuchung auftreten, die Nachher-Messung beeinflussen. Solche Effekte lassen sich mit Solomons Vier-Gruppen-Design kontrollieren (vgl. Campbell und Stanley 1966). Hier wird bei zwei randomisierten Gruppen (mit/ohne treatment) eine Vorher-Nachher-Messung durchgeführt, bei zwei weiteren randomisierten Gruppen (mit/ohne treatment) wird lediglich eine Nachher-Messung durchgeführt (Abbildung 2.3).

Abbildung 2.2: Solomons Vier-Gruppen-Design

		t_1	t_2	t_3	
Gruppe 1	R|	O	T	O	Vorher-Nachher-Messung
Gruppe 2	R|	O		O	
Gruppe 3	R|		T	O	Nachher-Messung
Gruppe 4	R|			O	

O: Beobachtung, T: treatment, R: Randomisierung

Bisher wurde von randomisierten Gruppen ausgegangen. Eine alternative Strategie zur Ausschaltung von systematischen Unterschieden zwischen Kontroll- und Experimentalgruppe ist die Parallelisierung (*matching*). Zwei Methoden werden unterschieden. Bei der Parallelisierung von Gruppen (*matched groups*) wird versucht, dieselbe Randverteilung bestimmter Merkmale in beiden Gruppen herzustellen, z. B. Geschlecht und Bildung. Die Gruppen haben dann einen gleichen Anteil an Männern und Frau-

en und die gleiche Bildungsstruktur. Bei der Parallelisierung von Paaren (*matched pairs*) wird versucht, dass in beiden Gruppen Personen mit den gleichen Merkmals*kombinationen* - z. B. Bildung und Geschlecht - vorhanden sind. Einer Frau mit Hochschulabschluss in der Experimentalgruppe wird dann eine Frau mit Hochschulabschluss in der Kontrollgruppe zugeordnet. Die Merkmalskombinationen von Bildung und Geschlecht sind dann in den Gruppen gleich. *Matching*-Verfahren sind der Randomisierung im allgemeinen unterlegen. Es ist kaum möglich, die Gruppen nach mehr als zwei Merkmalskombinationen zu matchen. Selbst wenn die Kombination von Geschlecht und Bildung in den Gruppen identisch ist, heißt dies nicht, dass die Gruppen in anderen für die Untersuchungsfrage bedeutenden Merkmalen (Drittvariablen) gleich sind. Dies ist der entscheidende Vorteil der Zufallsaufteilung.

Experimentelle Designs finden sich häufig in der der Psychologie und Medizin, in letzterer vor allem zum Testen der Wirksamkeit neuer Medikamente. Den Mitgliedern der Versuchsgruppe wird hierbei das Medikament verabreicht, während die Mitglieder der Kontrollgruppe lediglich ein Placebo erhalten. Von einem *Blindversuch* spricht man, wenn die Probanden nicht wissen, ob sie zur Experimental- oder Kontrollgruppe gehören. Damit soll verhindert werden, dass die Versuchsteilnehmer aufgrund ihres Wissens um den Erhalt oder Nicht-Erhalt der Behandlung eine unterschiedliche Reaktion zeigen. *Doppelblindversuche* liegen vor, wenn weder die Probanden noch der Versuchsleiter wissen, wer zur Experimental- bzw. Kontrollgruppe gehört. Hiermit sollen unbewusste Beeinflussungen durch den Versuchsleiter ausgeschaltet werden. Stellt man bei den Mitgliedern der Experimentalgruppe eine Verbesserung des Gesundheitszustands fest, bei den Mitgliedern der Kontrollgruppe jedoch nicht, so kann dies bei randomisierten Gruppen auf die Wirkung des zuvor verabreichten Medikaments zurückgeführt werden.

Experimentelle Anordnungen zeichnen sich durch die Randomisierung und das kontrollierte Setzen der unabhängigen Variable aus. Aufgrund von Randomisierung und zeitlicher Abfolge von Stimulus (unabhängiger Variable) und Wirkung (abhängige Variable) ermöglichen Experimente die Prüfung kausaler Hypothesen. Dies ist der entscheidende Vorteil im Vergleich zu den nachfolgend diskutierten Ex-post-facto-Andordnungen. Ein Nachteil experimenteller Untersuchungsformen besteht allerdings in der Schwierigkeit der Verallgemeinerung der Ergebnisse, d. h. der *externen Validität*. Vor allem deshalb, weil bei Experimenten meistens andere –

einfachere – Bedingungen hergestellt werden, als sie in der Realität vorherrschen. Dies gilt in besonderem Umfang für Experimente in künstlichen Umgebungen (*Laborexperimente*).

In den Sozialwissenschaften sind experimentelle Untersuchungsanordnungen mit randomisierten Gruppen quantitativ eher von untergeordneter Bedeutung (vgl. exemplarisch Gschwend und Hooghe 2008). Sie scheinen aber in den vergangenen Jahrzehnten an Bedeutung gewonnen zu haben, wie eine Auswertung der Artikel der American Political Science Review nahelegt (vgl. Druckman et al. 2006). Die Dominanz nicht-experimenteller Untersuchungsanordnungen liegt darin begründet, dass der Stimulus häufig nicht vom Forscher gesetzt werden kann. Man denke z. B. an die Frage, ob Arbeitslosigkeit politische Apathie erzeugt. Hier lassen sich keine Gruppen bilden, deren Mitglieder nach dem Prinzip der Zufallsaufteilung Arbeitslosigkeit ausgesetzt werden (Experimentalgruppe) oder nicht (Kontrollgruppe), um anschließend politisches Verhalten zu messen. (Es wäre ethisch auch nicht vertretbar.)

Methodenexperimente werden häufiger durchgeführt. So wurden im Allbus 2006 mit dem Themenschwerpunkt „Einstellungen gegenüber ethnischen Gruppen in Deutschland“ bei der Erhebung von einigen Einstellungen gegenüber Ausländern (vgl. Abbildung 3.2, S. 49) die Privatheit der Befragungssituation variiert (vgl. Wasmer et al. 2007). Bei einem Teil der Befragten wurden die Einstellungen vom Interviewer in einer mündlichen Befragung mit Computer erhoben (computer assisted personal interview, CAPI). Ein anderer Teil der Befragten füllte die Antworten selbst am Computer aus (computer assisted self interview, CASI). Die Zuteilung zur CAPI bzw. CASI-Gruppe erfolgte zufällig. Es handelt sich also um randomisierte Gruppen. Das Selbstausfüllen ist weniger „öffentlich“, weil die Einstellungen nicht dem Interviewer mitgeteilt werden müssen. Man kann daher erwarten, dass die CASI-Befragten eine geringere Tendenz aufweisen, ihre Antworten an dem auszurichten, was sie für sozial erwünscht halten (vgl. Kapitel 4.1). Mit dem ALLBUS 2006 kann dies kontrolliert werden, weil eine Skala zur Messung sozialer Erwünschtheit und die von den Befragten als sozial erwünscht angesehenen Positionen zu den vier Aussagen erhoben wurden.

Die **Ex-post-facto-Anordnung** ist die in den Sozialwissenschaften am häufigsten vorkommende Untersuchungsanordnung. Dabei kann es sich um eine Befragung, eine Beobachtung oder eine Inhaltsanalyse handeln (siehe

Kapitel 4). Die Untersuchungseinheiten – meist sind es Teilnehmer einer Befragung – werden erst *im Nachhinein* („ex post"), nämlich *bei der Datenauswertung*, in Experimental- und Kontrollgruppe unterteilt.

Untersucht man mit einer Umfrage, ob Arbeitslosigkeit die Wahl extremer Parteien begünstigt, so würde man die Stichprobe bei der Auswertung in Arbeitslose und Nicht-Arbeitslose aufteilen und für beide Gruppen das Wahlverhalten (gemessen z. B. durch die Wahlsonntagsfrage) ermitteln. Zeigt sich, dass Arbeitslose in stärkerem Umfang extreme Parteien wählen als Nicht-Arbeitslose, so heißt dies allerdings noch nicht, dass Arbeitslosigkeit politisch extremes Wahlverhalten *verursacht*, also ein kausaler Zusammenhang vorliegt. Warum?

In Experimenten können beobachtete Unterschiede auf die unabhängige Variable (den Stimulus) zurückgeführt werden, weil es sich um randomisierte Gruppen handelt und die unabhängige Variable der abhängigen Variable zeitlich vorgelagert ist. Randomisierte Gruppen unterscheiden sich bei Experimenten in der unabhängigen Variable (dem Stimulus) und möglicherweise in der abhängigen Variable (der Wirkung); sie unterscheiden sich jedoch nicht im Hinblick auf andere „dritte" Variablen, da die Zufallsaufteilung der Probanden eine gleiche Verteilung der Eigenschaften in Experimental- und Kontrollgruppe sicherstellt. Da die Gruppen sich nur in der unabhängigen Variable (Stimulus) unterscheiden, scheiden andere Faktoren als Ursache der Unterschiede in der abhängigen Variable aus. Experimente weisen daher eine hohe *interne Validität* auf. Bei Ex-post-facto-Anordnungen ist das anders. Die Gruppen sind nicht randomisiert, weshalb auch nicht mit Sicherheit gesagt werden kann, ob die Wirkung (Wahl extremer Parteien) auf die vermutete Ursache (Arbeitslosigkeit) zurückzuführen ist, oder ob sich die „Experimentalgruppe" (Arbeitslose) in anderen Merkmalen (z. B. Schulbildung) von der „Kontrollgruppe" (Nicht-Arbeitslose) unterscheidet, die ebenfalls einen Einfluss auf die die abhängige Variable (Wahl extremer Parteien) ausüben. Auch die zeitliche Reihenfolge der Variablen ist in Ex-post-facto-Anordnungen (zu Ausnahmen siehe Abschnitt 2.4) häufig unklar. Im Beispiel kann man davon ausgehen, dass Arbeitslosigkeit möglicherweise das Wahlverhalten beeinflusst, während die umgekehrte Wirkungsrichtung unplausibel ist.

Drittvariablenkontrolle

Wird bei einem Ex-post-facto-Design ein statistischer Zusammenhang (*Korrelation*) zwischen zwei Merkmalen X und Y festgestellt, so muss deshalb kontrolliert werden, ob die Ausprägung von Y tatsächlich auf X zurückgeführt werden kann (d. h. ein kausaler Einfluss von X naheliegt) oder ob alternative Erklärungen für Y existieren. Der Einfluss „dritter" (alternativer) Merkmale Z muss also geprüft werden. Für dieses Verfahren hat sich der Begriff Drittfaktor- oder Drittvariablenkontrolle eingebürgert. Mit der Drittvariablenkontrolle soll also verhindert werden, dass wir eine korrelative Beziehung (einen statistischen Zusammenhang) für eine kausale Beziehung halten. Dabei muss man sich im klaren sein, dass Kausalität in Ex-post-facto-Designs nicht empirisch „bewiesen" werden kann. Allerdings können die zeitliche Abfolge der Variablen und die Kontrolle von Drittvariablen eine kausale Interpretation eines statistischen Zusammenhangs nahelegen. Die Kontrolle von Drittvariablen setzt voraus, dass Hypothesen über den Einfluss dritter Merkmale vorhanden sind und auch entsprechende Daten zur Verfügung stehen. Letzteres stellt insbesondere bei Sekundärdatenanalysen ein Problem dar. Es kann also nicht ausgeschlossen werden, dass relevante Drittvariablen nicht kontrolliert wurden. Allerdings können kausale Hypothesen abgelehnt werden, etwa dann, wenn eine Korrelation bei Kontrolle einer Drittvariablen verschwindet (*Scheinkorrelation* oder *Scheinkausalität*).

Technisch wird eine Drittvariablenkontrolle bei Merkmalen mit wenigen Ausprägungen ausgeführt, indem getrennt für jede Ausprägung der Drittvariablen Z der Zusammenhang zwischen X und Y ermittelt wird. Die Ausprägung der Drittvariablen wird dadurch konstant gehalten. In Tabelle 2.1 ist der Zusammenhang zwischen der Schulbildung X und einer geringfügigen Beschäftigung Y (vgl. Bäcker 2007) wiedergegeben. Formal höher Gebildete sind prozentual in geringerem Umfang in Mini-/Midi-Jobs tätig, wie man im oberen Teil der Tabelle sehen kann. Der Unterschied beträgt 5 Prozentpunkte. Splittet man die Tabelle nach dem Geschlecht (Z) auf, dann ändern sich allerdings die Zusammenhänge. Bei Männern (Z_1) übt die Schulbildung (X) keinen Einfluss auf die Beschäftigungsform (Y) aus. Lediglich 3 % der Männer sind (unabhängig von der Schulbildung) geringfügig erwerbstätig. Bei Frauen (Z_2) sieht das Bild ganz anders aus. Ein Viertel (!) der Frauen mit niedriger Schulbildung (maximal Hauptschulabschluss) sind geringfügig beschäftigt, während die geringfügige Be-

schäftigung bei Frauen mit mittlerer und höherer Bildung einen deutlich geringeren Stellenwert einnimmt (12 %) (aber immer noch vier Mal höher ist als bei Männern). Bei Männern gibt es demnach keinen Zusammenhang zwischen X und Y, bei Frauen einen starken.

Tabelle 2.1: Schulbildung und geringfügige Beschäftigung 2006

			X	
			Schulbildung	
		Mini-/Midi-Job?	niedrig	mittel/hoch
Gesamt	Y	nein	88 %	93 %
		ja	12 %	7 %
			100 %	100 %
		Befragte	2.780	6.692
Z_1 = Männer	Y	nein	97 %	97 %
		ja	3 %	3 %
			100 %	100 %
		Befragte	1.648	3.370
Z_2 = Frauen	Y	nein	74 %	88 %
		ja	26 %	12 %
			100 %	100 %
		Befragte	1.132	3.322

Datengrundlage: Sozio-ökonomisches Panel (gewichtet), Welle W

In diesem Beispiel handelt es sich um eine *Interaktion*, die immer dann vorliegt, wenn sich ein Zusammenhang (Schulbildung und geringfügige Beschäftigung) in Abhängigkeit von der Ausprägung einer dritten Variablen (Geschlecht) ändert. Außer der Interaktion können weitere Effekte auftreten, die hier nur idealtypisch beschrieben werden können (vgl. Abbildung 2.4, S. 30). Der Zusammenhang zwischen X und Y kann durch eine Drittvariable Z bedingt sein, die sowohl X als auch Y beeinflusst. Ein triviales Beispiel wäre ein positiver Zusammenhang zwischen der Körpergröße (X) und den Leistungen (Y) von Schülern, wie sie etwa mit den PISA-Untersuchungen erhoben werden. Die Körpergröße ist nicht ursächlich für die besseren Leistungen. Vielmehr erklärt das Alter der Schüler (Drittvariable Z) sowohl deren Körpergröße als auch deren Leistungen (längere Schulbildung). Nach Kontrolle des Alters müsste der Zusammenhang zwischen der Körpergröße X und den Leistungen Y verschwinden (Abbildung 2.3). Man spricht hier von einer *Scheinkorrelation* oder bes-

ser von einer *Scheinkausalität*, denn die Korrelation zwischen X und Y besteht (ohne Kontrolle von Z) ja tatsächlich.

Abbildung 2.3: Scheinkausalität

Größe (X) → + → Leistung (Y)	Alter (Z) → + → Größe (X); Alter (Z) → + → Leistung (Y)
ohne Kontrolle von Z	mit Kontrolle von Z

Möglich ist auch, dass X Z beeinflusst und Z wiederum Y, X also einen *indirekten Effekt* auf Y ausübt (X → Z → Y). Ein Beispiel ist der im sozialpsychologischen Modell des Wahlverhaltens postulierte Einfluss der als langfristig stabil konzeptualisierten Parteiidentifikation (PI) (vgl. Campbell et al. 1980). Die Parteiidentifikation dient als Filter der Wahrnehmung der Kandidaten (K). Kandidaten der präferierten Partei werden positiver wahrgenommen als Kandidaten anderer Parteien. Die Kandidatenpräferenz beeinflusst die Stimmabgabe (Wahl). Liegt ausschließlich ein indirekter Effekt vor (PI → K → Wahl), dann verschwindet auch hier der ursprüngliche Zusammenhang zwischen der Parteiidentifikation und der Wahlentscheidung nach Kontrolle der Kandidatenpräferenz. Nach dem sozialpsychologischen Modell übt die Parteiidentifikation zudem einen direkten Effekt auf die Wahlentscheidung aus (*direkter und indirekter Effekt*). Der Unterschied zwischen einem indirekten Effekt und einer Scheinkorrelation lässt sich nur über die zeitliche Abfolge von X und Z klären. Bei einer Scheinkorrelation ist Z *X und Y* kausal vorgelagert, bei einem indirekten Effekt ist Z nur Y kausal vorgelagert. Von *Multikausalität* ist die Rede, wenn sowohl X als auch Z einen eigenständigen Einfluss auf die abhängige Variable Y ausüben. Zwei zentrale Determinanten der Lohnhöhe sind die schulische/berufliche Ausbildung und die Berufserfahrung. Schließlich kann ein Zusammenhang erst bei Kontrolle einer Drittvariablen auftauchen (*scheinbare Nichtkorrelation*, nicht abgebildet). Die Kontrolle von Drittvariablen ist demnach auch sinnvoll, wenn keine Korrelation festgestellt wurde. Zur Kontrolle von Drittvariablen werden in der Regel multivariate Verfahren eingesetzt. Multivariate Verfahren sind - vereinfacht

ausgedrückt - statistische Methoden zur Analysen von Zusammenhängen zwischen mehr als zwei Variablen. Indirekte Effekte lassen sich mit Pfadmodellen quantifizieren (vgl. Reinecke 2005). Einen Überblick über die statistische Kontrolle von Drittvariablen bieten Agresti und Finlay (2008, Kapitel 10) und Benninghaus (2005).

Abbildung 2.4: Beziehungen zwischen drei Variablen

Scheinkausalität: $Z \rightarrow X$, $Z \rightarrow Y$

Indirekter Effekt: $X \rightarrow Z \rightarrow Y$

Direkter & indirekter Effekt: $X \rightarrow Y$, $X \rightarrow Z \rightarrow Y$

Interaktion: $Z \downarrow (X \rightarrow Y)$

Multikausalität: $X \rightarrow Y$, $Z \rightarrow Y$

Neben der Feststellung der Kausalität existiert in Ex-post-facto-Anordnungen ein weiteres Problem. Die Größe von „Experimental-“ und „Kontrollgruppe“ kann nicht so gezielt gesteuert werden, da im Gegensatz zu Experimenten die Gruppenaufteilung erst bei der Datenauswertung erfolgt. Aus diesem Grunde kommt es in Ex-post-facto-Anordnungen vor, dass Merkmalsausprägungen, die untersucht werden sollen, zu selten auftreten. Eine Untersuchung der Wähler der Republikaner mit dem ALLBUS 1998 scheitert schlicht daran, dass lediglich 53 der 3.432 Befragten eine Republikaner-Wahlabsicht angaben. Dieses Problem kann allerdings durch größere Stichproben oder geschichtete Auswahlverfahren (siehe Kapitel 9) gelöst werden.

2.4 Zeitdimension

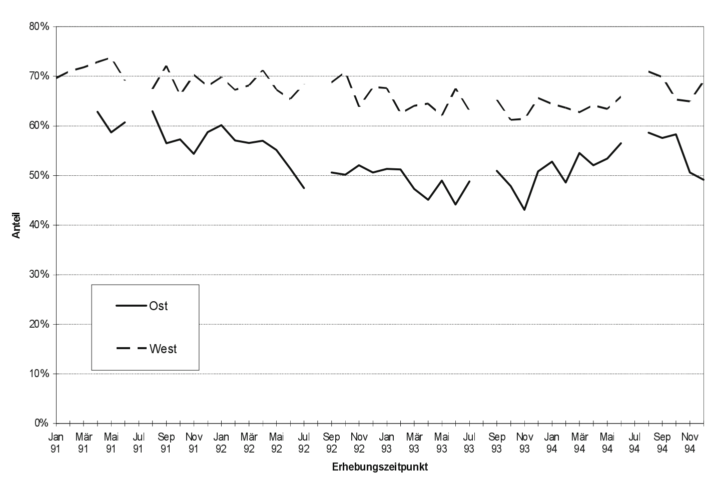
Hinsichtlich der Zeitdimension lassen sich Forschungsdesigns danach unterscheiden, ob die Erhebung zu einem Zeitpunkt (*Querschnittdesign*) oder mehreren Zeitpunkten (*Längsschnittdesign*) stattfindet (vgl. Bijleveld und van der Kamp 1998). Zu den Längsschnittdesigns zählt das *Trenddesign* und das *Paneldesign*.

- Querschnittdesign
- Trenddesign
- Paneldesign

Bei einem **Querschnittdesign** erfolgt die Datenerhebung zu einem einzigen Zeitpunkt bzw. in einem kurzen Zeitintervall. Mit Querschnittdaten sind die in Kapitel 2.3 erwähnten Probleme der Überprüfung kausaler Hypothesen verbunden.

Ein **Trenddesign** liegt vor, wenn *dieselben Merkmale* zu *verschiedenen Zeitpunkten* an *unterschiedlichen Stichproben* gemessen werden. Eine Trendstudie lässt sich damit als Abfolge von mehreren Querschnittstudien auffassen. Um einen Trend feststellen zu können, müssen die Stichproben repräsentativ für die gleiche Grundgesamtheit sein (vgl. Kapitel 9). Bei den ALLBUS-Studien, den Politbarometern der Forschungsgruppe Wahlen, und den Eurobarometerumfragen handelt es sich um Trendstudien (siehe dazu S. 37). Ein Beispiel für eine Trendauswertung der Politbarometer-Umfragen der Forschungsgruppe Wahlen ist in Abbildung 2.5 zu sehen. Angegeben ist die Entwicklung der Parteiidentifikation zwischen Januar 1991 und November 1994 getrennt für ost- und westdeutsche Befragte. Auf der x-Achse ist die Zeit abgetragen, auf der y-Achse der jeweilige Anteil der Befragten, die sich mit einer Partei identifizieren. Die Linienzüge sind in den Monaten unterbrochen, in denen keine Politbarometer-Umfragen durchgeführt wurden. Im Mai 1991 gaben mehr als 70 % der westdeutschen Befragten an, sich mit einer Partei zu identifizieren, während der Anteil ostdeutscher Befragter mit Parteiidentifikation mit etwas mehr als 60 % knapp 10 Prozentpunkte niedriger liegt. Bis November 1993 nimmt der Anteil der Befragten mit Parteiidentifikation im Osten und Westen ab, um dann bis Herbst 1994 wieder anzusteigen. Man sieht, dass die Schwankungen des Anteils der Personen mit Parteiidentifikation im Osten der Republik deutlicher ausfallen als im Westen.

Abbildung 2.5: Parteiidentifikation zwischen 1991 und 1994



Quelle: Gehring und Winkler (1997), monatlich n≈1000

Mit den Politbarometerdaten kann nicht untersucht werden, inwieweit mit der Frage zur Parteiidentifikation tatsächlich langfristig stabile Bindungen an eine Partei erfasst werden, wie es das Konzept vorsieht. Über individuelle Veränderungen der Parteiidentifikation im Zeitverlauf können keine Aussagen getroffen werden, da die monatlichen Umfragen auf unterschiedlichen Stichproben beruhen. Auf der Basis von Trenddaten lassen sich also Veränderungen im Aggregat (hier West bzw. Ost), die so genannten *Nettoveränderungen*, feststellen, nicht aber Veränderungen bei einzelnen Untersuchungseinheiten (*Bruttoveränderungen*, vgl. Engel und Reinecke 1994, 6).

Unter einem **Paneldesign** versteht man eine Erhebung *derselben Merkmale* zu *verschiedenen Zeitpunkten* an *denselben Untersuchungseinheiten*. Die einzelnen Befragungszeitpunkte werden als Wellen bezeichnet. In der Regel handelt es sich nur um wenige Wellen. Eine Ausnahme stellt das Sozio-ökonomische Panel dar, das bereits seit 1984 einmal jährlich durchgeführt wird (vgl. zum Untersuchungsdesign S. 37).

In Abbildung 2.6 ist das Design einer Studie von Paul F. Lazarsfeld, Bernard Berelson und Hazel Gaudet zur US-amerikanischen Präsidentschaftswahl 1940 dargestellt, die in Erie-County, einem Kreis in Ohio, durchgeführt wurde (vgl. Lazarsfeld et al. 1968). Für die Demokraten trat der Amtsinhaber Franklin D. Roosevelt an, der die Wahl gegen den republikanischen Herausforderer Wendell L. Willkie für sich entscheiden konnte. Im Mai 1940 wurden 3.000 Personen befragt. Von diesen Befragten wurden 600 Personen für das Hauptpanel (HP) ausgewählt und zu sechs weiteren Zeitpunkten befragt. Beim Hauptpanel handelt es sich also um ein siebenwelliges Panel. Zusätzlich wurden aus der Gesamtstichprobe drei weitere Stichproben à 600 Personen gezogen (Stichproben A, B und C). Diese Personen wurden aus methodischen Gründen – zur Kontrolle von Paneleffekten – außer im Mai noch zu je einem weiteren Zeitpunkt befragt, die Teilnehmer der Stichprobe A z. B. im Juli. Für Juli, August und Oktober existieren somit Ergebnisse einer Vergleichsstichprobe.

Abbildung 2.6: Forschungsdesign der Untersuchung „The People's Choice"

Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.
Gesamtbefragung n=3000	Hauptpanel, n=600					
	HP	HP	HP	HP	HP	HP
	Kontrollstichproben					
		A n=600	B n=600		C n=600	

Vgl. Lazarsfeld et al. (1968, S. 4)

In Tabelle 2.2 wurde das in der Nachwahlbefragung (im November) angegebene Stimmverhalten mit der im Oktober angegebenen Wahlabsicht gekreuzt. (Wie man sieht, lagen die Republikaner in Erie-County vorne.) In der Summenspalte und Summenzeile lassen sich die Veränderungen im Aggregat, die Nettoveränderungen, beobachten: Im November stimmten von allen 483 Personen 48 % (232) für die Republikaner, während dies im Oktober 47 % (229) beabsichtigt hatten. Durch die Veränderungen im Aggregat wird das Ausmaß des Wandels auf der individuellen Ebene unterschätzt: von den 483 Personen stimmen 418 Personen (215 + 144 + 59) genau so, wie sie es im Oktober beabsichtigten; dies entspricht 87 %. 13 %

der Befragten änderten zwischen den beiden Zeitpunkten ihre Präferenz (vgl. Lazarsfeld et al. 1968, xxiii); diese individuellen Veränderungen werden auch *turnover* genannt. So gingen 10 Personen, die im Oktober die Wahl der Republikaner beabsichtigten, nicht zur Wahl. 11 Personen wechselten die Parteipräferenz: 7 von den Demokraten zu den Republikanern und 4 umgekehrt. Mit Paneldaten lassen sich also sowohl die Veränderungen auf der Aggregatebene (Nettoveränderungen) als auch die individuellen Veränderungen (Bruttoveränderungen) untersuchen.

Tabelle 2.2: Wahlabsicht und Stimmabgabe in Erie-County bei den Präsidentschaftswahlen in den USA, 1940

Stimmabgabe (November)	Wahlabsicht (Oktober) Rep.	Dem.	N. w.	w. n.	Summe
Republikaner	*215*	7	6	4	232
Demokraten	4	*144*	0	12	160
Nichtwahl	10	16	*59*	6	91
Summe	229	167	65	22	483

Rep. = Republikaner, Dem. = Demokraten, N. w. = Nichtwahl, w. n.= weiß nicht Quelle: Lazarsfeld et al. (1968, S. xxiii)

Von Panelanalysen spricht man nur dann, wenn die Veränderungen der Merkmalsausprägungen von Untersuchungseinheiten im Zeitverlauf betrachtet werden. Berücksichtigt man lediglich Aggregatveränderungen, dann handelt es sich um eine Trendanalyse auf der Basis von Paneldaten. Wertet man wie in Tabelle 2.1 lediglich eine Welle eines Panels aus, dann ist es eine Querschnittanalyse. Aufgrund der durch die unterschiedlichen Messzeitpunkte klaren zeitlichen Abfolge der Variablen eignen sich Panelanalysen besser zur Überprüfung kausaler Zusammenhänge als Ex-post-facto-Designs. Auch hier ist es jedoch notwendig, alternative Erklärungen durch die Kontrolle von Drittvariablen auszuschließen.

Panelstudien sind mit besonderen methodischen Problemen konfrontiert: der *Panelmortalität* und *Paneleffekten*. Unter **Panelmortalität** wird die Tatsache verstanden, dass nicht alle Befragten der ersten Untersuchung auch bei den folgenden Untersuchungen wieder befragt werden können, sei es, weil sie umgezogen oder aus anderen Gründen nicht mehr erreichbar sind, die wiederholte Teilnahme verweigern oder zwischenzeitlich verstorben sind. Auf diese Weise verringert sich der Bestand eines Panels stetig.

Üblicherweise kann man davon ausgehen, dass der Bestand mit jeder Folgeuntersuchung abnimmt, wobei die zweite Untersuchung normalerweise die höchsten Bestandsverluste aufweist. Dies kann dazu führen, dass Fragestellungen nicht mehr untersucht werden können, weil einfach zu wenige Personen eine interessierende Merkmalsausprägung aufweisen. Um der Panelmortalität entgegenzuwirken, können besondere Maßnahmen ergriffen werden, die als *Panelpflege* bezeichnet werden. Die Panelpflege dient dazu, die Befragten zur weiteren Teilnahme zu motivieren und den Kontakt zu Befragten, die den Wohnsitz wechseln, nicht zu verlieren. Die Teilnehmer des Sozio-ökonomischen Panels erhalten ein kleines Geschenk, nehmen an einer bundesweiten Lotterie teil und werden über zentrale Ergebnisse der Umfrage informiert (vgl. Haisken-DeNew und Frick 2005, 27).

Als **Paneleffekte** werden Auswirkungen der wiederholten Befragung auf die Meinungen und Einstellungen der Panelteilnehmer bezeichnet. Ein Problem tritt dann auf, wenn die Zeit zwischen den Befragungen sehr kurz ist und der Befragte sich an seine vergangenen Angaben erinnert. Dies kann zur Konsequenz haben, dass die Befragten versuchen, möglichst konsistent zu antworten. Die Stabilität der Antworten würde dann überschätzt werden. Wiederholte Interviews können auch dazu führen, dass den Befragten ihre eigenen Ansichten und Meinungen bewusster werden, weil sie sich häufiger mit den Befragungsthemen beschäftigen. Paneleffekte lassen sich kontrollieren, in dem zeitgleich zu einer Panelwelle eine Kontrollstichprobe mit denselben Messinstrumenten untersucht wird. Die Meinungen und Einstellungen der Panelteilnehmer können dann mit denen der Teilnehmer der Kontrollstichprobe verglichen werden. Genau dies war der Sinn der Kontrollstichproben A, B und C in der Studie von Lazarsfeld et al. (1968) (vgl. Abbildung 2.6). Es zeigte sich, dass die Panelteilnehmer ihre Wahlentscheidung früher trafen als die Teilnehmer der Kontrollstichproben (vgl. Lazarsfeld et al. 1968, xv).

Den Designs entsprechend kann man zwischen *Querschnittdaten*, *Zeitreihendaten* und *Paneldaten* unterscheiden. Eine besondere Form von Daten stellen zudem *Verlaufsdaten* dar. Um Zeitreihendaten handelt es sich beispielsweise bei der Entwicklung des Anteils der Personen, die eine Parteiidentifikation zwischen 1991 und 1994 in der Bundesrepublik aufweisen. Zeitreihendaten beziehen sich auf eine Einheit (Bundesrepublik) zu mehreren Zeitpunkten (1991-1994). Paneldaten sind Daten, die – wie erwähnt – Informationen über individuelle Veränderungen beinhalten. Verlaufsdaten geben zusätzlich Auskunft über die Länge eines Zeitintervalls

bis zum Eintritt eines Ereignisses, z. B. die Dauer bis zur Änderung der Parteiidentifikation oder die Dauer bis zur Aufnahme einer Beschäftigung nach Arbeitslosigkeit. Verlaufsdaten werden daher auch als Ereignisdaten bezeichnet. Aus den Daten des Sozio-ökonomischen Panels können Verlaufdaten generiert werden. Die Dauer der Parteiidentifikation, die im SOEP jährlich erfasst wird, bestimmt sich aus dem Zeitraum zwischen der erstmaligen Angabe der Identifikation mit einer bestimmten Partei bis zur eventuellen Aufgabe der Identifikation. Noch exaktere Angaben liefert das SOEP unter anderem für die Erwerbsbiographie. Die Befragten füllen ein so genanntes Kalendarium aus, indem für jeden Monat des der Befragung vorangegangenen Kalenderjahres der Erwerbsstatus rückwirkend erfasst wird. Die Dauer der Arbeitslosigkeit kann daher monatsgenau ermittelt werden. Mit einer Ereignisdatenanalyse (vgl. Blossfeld et al. 2007) der SOEP-Daten zeigen Schmitt-Beck et al. (2006), dass die Parteiidentifikation zwischen 1984 und 2001 nur bei einem kleinen Teil der Wähler über die gesamte Periode stabil war.

Um Paneldaten oder Ereignisdaten zu erheben, muss man nicht zwangsläufig dieselben Personen mehrmals befragen. Auch mit einem Querschnittdesign kann man Panel- und Ereignisdaten gewinnen; durch Fragen, die sich auf die Vergangenheit beziehen (*Retrospektivfragen* bzw. *Recall-Fragen*). Erhebt man Angaben zum aktuellen Wahlverhalten und dem Wahlverhalten bei der vorangegangenen Wahl, dann hat man Paneldaten (siehe unten) erhoben, mit denen Wechselwahlverhalten untersucht werden kann. Allerdings stellen retrospektive Fragen hohe Ansprüche an die Erinnerungsfähigkeit des Befragten. Diese wird umso besser sein, je kürzer die Ereignisse zurückliegen und je wichtiger diese für den Befragten sind. Daten zur Biographie müssten beispielsweise relativ präsent sein und daher auch zuverlässig abgefragt werden können. Einstellungen oder Meinungen können mit Retrospektivfragen kaum zuverlässig erfasst werden. Zudem besteht die Gefahr, dass Befragte Widersprüche zwischen vergangenen und gegenwärtigen Einstellungen aufzulösen versuchen, in dem sie vergangene Einstellungen den aktuellen „anpassen“. Auch bei Angaben zum Wahlverhalten bei einer früheren Wahl stellt sich also die Frage, ob diese korrekt sind (vgl. Schoen 2000).

Forschungsdesigns einiger sozialwissenschaftlicher Erhebungen

ALLBUS Die *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften* – ALLBUS – wird von der GESIS, Arbeitsbereich Dauerbeobachtung (früher: Zentrum für Umfragen, Methoden und Analysen, ZUMA), seit 1980 durchgeführt. In zweijährigen Abständen werden ca. 3.000 Befragte zu sozialwissenschaftlich interessanten Themen befragt. Es handelt sich hier um eine Querschnittbefragung, die sich für Trendstudien eignet, da ein Teil der Fragen bereits mehrmals erhoben wurde. Eine Ausnahme vom Befragungsrhythmus bildete die 1991 erhobene ALLBUS-Basisumfrage, wo erstmals Personen in Ostdeutschland mit in die Befragung aufgenommen wurden (vgl. Alba et al. 2000; Braun und Mohler 1998, 1994).

Eurobarometer Das Eurobarometer ist eine im Auftrag der Kommission der Europäischen Union seit 1974 zweimal jährlich durchgeführte Befragung in den Ländern der Europäischen Union. Ebenso wie beim Politbarometer und dem ALLBUS handelt es sich hier um Querschnittbefragungen, die für Trendstudien verwendet werden können. Das Eurobarometer eignet sich sehr gut für vergleichende Analysen, da in jedem Land mit weitgehend demselben Fragebogen gearbeitet wird. Mit Ausnahme Luxemburgs werden in jedem Land ca. 1.000 Personen befragt.

Politbarometer Das Politbarometer wird im Auftrag des ZDF seit 1977 von der Forschungsgruppe Wahlen in Mannheim durchgeführt. Es handelt sich um eine monatliche Querschnittbefragung von ca. 1.250 Personen. Einzelne Ergebnisse werden einmal im Monat im ZDF verbreitet und sind über Internet abrufbar. Die erhobenen Daten werden an das *Zentralarchiv für empirische Sozialforschung* (ZA) in Köln weitergegeben, wo sie aufbereitet und für wissenschaftliche Zwecke zur Verfügung gestellt werden. Wegen des immer gleichen Designs und der Verwendung derselben Fragen eignet sich das Politbarometer ausgezeichnet für Trendstudien.

SOEP Das Sozio-ökonomische Panel (SOEP) ist die umfangreichste Panelstudie in der Bundesrepublik. Es handelt sich um eine Haushaltsstichprobe, bei der alle Mitglieder eines Haushaltes befragt werden, die im Befragungsjahr mindestens 17 Jahre alt sind oder werden. Beginnend 1984, wurden ca. 12.000 Personen aus 6.000 Haushalten (Stichprobe A „Deutsche“ und B „Ausländer“) einmal jährlich

befragt, seit 1990 auch in Ostdeutschland (Stichprobe C). Eine wesentlicher Vorteil des SOEP ist die große Ausländerstichprobe (Stichprobe B), die detaillierte Analysen ermöglicht, und die Erfassung der neueren Formen von Zuwanderung (Stichproben D1/D2). Inzwischen wurde das SOEP um weitere Stichproben ergänzt (vgl. Tabelle 2.3). Durch die Auffrischung des Panels wird die Panelmortalität aufgefangen. Die Auffrischung dient zudem zur Sicherstellung der „Repräsentativität“ des Panels, da die ursprünglichen Stichproben zwar die Bevölkerung im Jahre 1984 repräsentieren, der Veränderung der Bevölkerungsstruktur aber nicht Rechnung tragen.

Tabelle 2.3: Stichprobenstruktur des Sozio-ökonomischen Panels

Stichprobe		Start in	Haushalte
A	Westdeutsche	1984	4.528
B	Ausländer	1984	1.393
C	Ostdeutsche	1990	2.179
D1/D2	Zuwanderer	1994/1995	522
E	Erneuerungsstichprobe	1998	1.067
F	Innovationsstichprobe	2000	6.052
G	Hocheinkommensstichprobe	2002	1.224
H	Auffrischungsstichprobe	2006	1.506

Im SOEP werden unter anderem detaillierte Angaben zur Demographie der Befragten, zu deren sozialer und ökonomischer Situation im Haushaltskontext, zu deren Erwerbssituation und Einkommensverläufen erhoben. Neben einer Reihe von Indikatoren, die für soziologische und ökonomische Fragestellungen relevant sind, enthält das SOEP einige originär politikwissenschaftliche Merkmale, wie die bereits erwähnte Frage nach der Parteiidentifikation. Mit der Durchführung des SOEP ist eine Projektgruppe am Deutschen Institut für Wirtschaftsforschung (DIW) in Berlin beauftragt (vgl. Hanefeld 1987; Haisken-DeNew und Frick 2005).

Adressen der genannten Institutionen

CSDM Center for Survey Design and Methodology der GESIS (früher: Zentrum für Umfragen, Methoden und Analysen, ZUMA), B 2,1, 68159 Mannheim; Postanschrift: Postfach 122155, 68072 Mannheim;

Tel.: 0621/1246-0, Fax: -100.
Internet: `http://www.gesis.org/dienstleistungen/methoden/`

DIW Deutsches Institut für Wirtschaftsforschung: Königin-Luise-Str. 5, 14195 Berlin; Tel.: 030/89789-0, Fax: 030/89789-200.
Internet: `http://www.diw-berlin.de/`

FDZ Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit - Forschungsdatenzentrum -, Regensburger Str. 104, 90478 Nürnberg; Tel.: 0911/179-1752, Fax: 0911/179-1728.
Internet: `http://www.fdz.iab.de/`

FGW Forschungsgruppe Wahlen: N 7, 13-15, 68161 Mannheim; Postanschrift: Postfach 101121, 68011 Mannheim; Tel.: 0621/ 1233-0, Fax: 0621/1233-199.
Internet: `http://www.forschungsgruppe.de/`

GESIS, Abteilung Datenarchiv und Datenanalyse (früher: Zentralarchiv für empirische Sozialforschung): Bachemer Str. 40, 50931 Köln; Postanschrift: Postfach 410960, 50869 Köln; Tel.: 0221/47694-0, Fax: -44.
Internet: `http://www.gesis.org/dienstleistungen/daten/`

StBA Statistisches Bundesamt: Gustav-Stresemann-Ring 11, 65180 Wiesbaden; Tel.: 0611/75-1, Fax: 0611/724000.
Internet: `http://www.destatis.de/`

Aufgaben zu Forschungsdesigns

1. Charakterisieren Sie den Unterschied zwischen Individual- und Aggregatdaten!
2. Sie haben vom Statistischen Bundesamt in Wiesbaden die Wahlergebnisse der Bundestagswahl 1994 und die Arbeitslosenquote für die Bundestagswahlkreise erhalten. In Ihrer Untersuchung stellen Sie nun einen Zusammenhang zwischen der Höhe der Arbeitslosenquote und dem Anteil der Stimmen für die Republikaner fest. Welchen Fehlschluss können Sie bei Analyse der Daten begehen und warum?
3. Worin unterscheiden sich die behandelten Längsschnittanalysen und welche Vor- bzw. Nachteile haben diese?
4. Im ALLBUS wurden die Einstellungen zur innerfamilialen Arbeitsteilung seit 1982 unter anderem mit der Aussage *„Es ist für alle Beteiligten viel besser, wenn der Mann voll im Berufsleben steht und die Frau zu Hause bleibt und sich um den Haushalt und die Kinder kümmert“* erfasst. In der folgenden Tabelle sind die Anteile der westdeutschen Befragten, die der Aussage zustimmten (stimme voll und ganz zu/stimme eher zu) bzw. sie ablehnten (lehne ab/lehne voll und ganz ab), wiedergegeben:

	1982	1992	1996	2000	2004
stimme zu	70 %	56 %	51 %	50 %	42 %
stimme nicht zu	30 %	44 %	49 %	50 %	58 %
	100 %	100 %	100 %	100 %	100 %
Anzahl der Befragten	2.910	2.325	2.326	2.425	1.936

   Beschreiben Sie die inhaltliche Aussage der Tabelle. Wie haben sich die Einstellungen im Zeitverlauf geändert? Handelt es sich um Querschnitt- und/oder um Längsschnittdaten? Begründen Sie Ihre Antwort!
5. Mit welchem Untersuchungsdesign kann man kausale Zusammenhänge feststellen?
6. Zählen die Volkszählungsdaten zu den Individual- oder Aggregatdaten?

3 Messen

3.1 Messen in der empirischen Sozialforschung

Wie wir in Kapitel 1 erläutert haben, entscheidet über die Aufrechterhaltung oder das Verwerfen einer Theorie oder einer Hypothese die Konfrontation mit der Realität. Das Messen spielt daher eine (wenn nicht sogar *die*) zentrale Rolle innerhalb der empirischen Sozialforschung. Bevor soziale Phänomene gemessen werden können, sind jedoch eine Reihe von Vorüberlegungen notwendig.

Ausgangspunkt einer Untersuchung sind Theorien und die in ihnen enthaltenen Hypothesen. Zunächst müssen die in den Theorien bzw. Hypothesen enthaltenen Begriffe präzise definiert sein. Bevor also ein Begriff wie „Rechtsextremismus" gemessen werden kann, muss geklärt werden, was darunter verstanden wird und welche Dimension(en) der Begriff beinhaltet. Anschließend muss der Begriff operationalisiert werden. Unter Operationalisierung werden alle Forschungsvorgänge („Operationen") verstanden, die notwendig sind, um festzustellen, „ob und in welchem Ausmaß der mit dem Begriff bezeichnete Sachverhalt in der Realität vorliegt" (Kromrey 2006, 189).

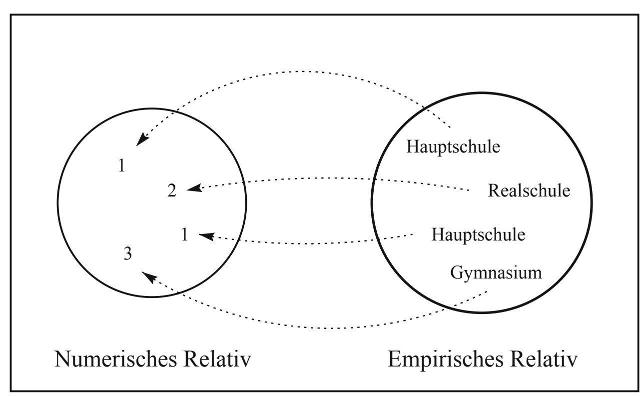
In Kapitel 1 wurden die CASMIN-Klassifikation und die Zahl der Schuljahre als mögliche Operationalisierungen von Bildung eingeführt. Bildung könnte genau so gut über einen Wissenstest erfasst werden. Die CASMIN-Skala erfasst schulische und berufliche Zertifikate, ein Wissenstest tatsächlich vorhandene Kompetenzen im getesteten Bereich. Solche Kompetenzmessungen werden im Rahmen der PISA-Untersuchungen inzwischen regelmäßig bei Schülerinnen und Schülern durchgeführt. Welche Operationalisierung angemessen ist, hängt von der Fragestellung der Untersuchung ab.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_3,

Was wird nun in den Sozialwissenschaften unter „Messen" verstanden? Wir begnügen uns im Folgenden mit einer „weichen" (und nicht ganz präzisen) Begriffsbestimmung des Messens, ohne auf die axiomatische Messtheorie zurückzugreifen (vgl. Kromrey 2006; Orth 1974).

Beim Messen geht es – wie Stevens (1946) definiert hat – um die **Zuordnung von Zahlen (oder Symbolen) zu Objekten nach bestimmten Regeln**. „Nach bestimmten Regeln" soll heißen, dass die Zuordnung so erfolgen muss, dass die Beziehungen zwischen den Zahlen die Beziehungen zwischen den Objekten **strukturtreu** widerspiegeln (nicht umgekehrt!). Die Beziehungen zwischen den Objekten werden empirisches Relativ genannt, die Beziehungen zwischen den Zahlen numerisches Relativ (Beziehungen = Relationen). Zum Messvorgang gehören damit drei Komponenten: Das **empirische Relativ**, das **numerische Relativ** und eine **Abbildungsvorschrift**, die eine korrekte (strukturtreue) Zuordnung der Zahlen zu den Eigenschaften von Objekten ermöglicht (vgl. Abbildung 3.1). Diese drei Komponenten bilden eine Skala.

Abbildung 3.1: Messen – Schematische Darstellung



Objekte haben in der Regel viele Eigenschaften, anhand derer sie in Beziehung gesetzt werden können. Bei Personen können dies z. B. das Geschlecht, die Bildung, das Einkommen, die Stärke des Politikinteresses oder die Wahlabsicht sein. Will man das Geschlecht messen, so könnte die

Abbildungsvorschrift lauten: Ordne den Merkmalsausprägungen männlich und weiblich die Zahlen 1 und 2 zu. Die konkrete Zuordnung ist beliebig: Männlich kann 1 sein und weiblich 2 oder umgekehrt. Die Zuordnung muss aber innerhalb einer Untersuchung konstant erfolgen. Über eine Ordnung macht die obige Abbildungsvorschrift keine Aussage, da im empirischen Relativ keine Ordnung vorliegt. Die Abbildungsvorschrift für das Politikinteresse würde dagegen lauten: Ordne die Zahlen so zu, dass die Rangfolge in der Stärke des Politikinteresses erhalten bleibt, also beispielsweise bei keinem Interesse 0, schwachem Interesse 1 und starkem Interesse 2. Hier ist eine Ordnung notwendig, um die Beziehungen im empirischen Relativ strukturtreu abzubilden. Das heißt: Die Abbildungsvorschrift ist **von der Art des Merkmals im empirischen Relativ abhängig** und bestimmt das Messniveau bzw. *Skalenniveau*.

3.2 Skalenniveaus

In Anlehnung an Stevens (1946) werden in den Sozialwissenschaften vier Skalenniveaus (auch: Messniveau) unterschieden:

1. Nominalskala
2. Ordinalskala
3. Intervallskala
4. Ratioskala

Außerdem ist die Berücksichtigung einer weiteren Skala sinnvoll:

5. Absolutskala

Das niedrigste Messniveau weist die Nominalskala auf, das höchste die Absolutskala. **Die höheren Skalen besitzen die Eigenschaften aller niedrigeren Skalen**. Nominal- und Ordinalskalen werden auch als *qualitative* oder *nicht-metrische* Skalen bezeichnet, Intervall-, Ratio- und Absolutskalen als *quantitative* oder *metrische* Skalen.

1. **Nominalskala**: Können die Ausprägungen von Merkmalen lediglich im Hinblick auf **Gleichheit** oder **Ungleichheit** unterschieden werden, dann liegt nominales Skalenniveau vor. Typische Beispiele hierfür sind das Geschlecht, die Parteipräferenz, die Haarfarbe oder die Religionszugehörigkeit. Welche Zahlen welcher Ausprägung zugeordnet werden, ist beliebig, solange für jede Merkmalsausprägung eine

eigene Zahl verwendet wird. Ob Männer mit 1 und Frauen mit 2 oder erstere mit 20 und letztere mit 17 bezeichnet werden, ist vollkommen unerheblich. Allerdings darf nur die Ungleichheit zwischen den Zahlen interpretiert werden. Die Aussage Frauen seien „doppelt so gut“ wie Männer, weil Frauen mit 2 und Männer „nur“ mit 1 kodiert wurden, ist sinnlos, da die Zahlenzuordnung beliebig ist.

2. **Ordinalskala**: Von ordinalem Skalenniveau spricht man, wenn die Merkmalsausprägungen zusätzlich zur Gleichheit/Ungleichheit noch eine **Reihenfolge** aufweisen. Bekannt ist *nur* die Reihenfolge; man weiß aber nicht, wie groß die Abstände zwischen den Merkmalsausprägungen sind. Es wurde bereits die Stärke des Politikinteresses genannt, andere Beispiele für ordinalskalierte Merkmale sind die Schulbildung und die Schulnoten. Die Reihenfolge der Merkmalsausprägungen muss sich in der Reihenfolge der Zahlen widerspiegeln. Die Zahlen müssen aber nicht unmittelbar aufeinander folgen, obwohl dies in der Regel – wie z. B. bei den Schulnoten – der Fall ist.
3. **Intervallskala**: Merkmale sind intervallskaliert, wenn deren Ausprägungen nicht nur eine Rangfolge (und damit auch Unterschiedlichkeit) aufweisen, sondern auch **Abstände** zwischen Ausprägungen sinnvoll interpretiert werden können . Typische Beispiele sind die Temperaturmessung in Celsius oder Fahrenheit und die Kalenderzeitrechnung. Die Abstände zwischen aufeinanderfolgenden Ausprägungen (die Intervalle) sind bei einer Intervallskala gleich groß (konstant). Der Altersunterschied zwischen einer Person, die 1930 geboren wurde und einer Person, die 1929 zur Welt kam, ist genauso groß wie die zwischen dem Geburtsjahrgang 1951 und 1950. Intervallskalen besitzen im Gegensatz zu den nachfolgend beschriebenen Ratioskalen aber keinen natürlichen Nullpunkt. Aus diesem Grund sind Verhältnisse zwischen den Zahlen auch nicht interpretierbar. Besonders deutlich wird dies an der Zeitrechnung: Als wir nach christlicher Zeitrechnung (gregorianischer Kalender) den Jahreswechsel 2008/2009 begangen haben, befanden wir uns nach jüdischer Zeitrechnung mitten im Jahr 5769 und nach islamischer Zeitrechnung im Jahr 1430. Das Jahr 0 existiert zwar bei allen drei Zeitrechnungen, es sind jedoch rein definitorische Festlegungen und keine „echten“ Nullpunkte. Der Beginn der Zeitrechnung ist im christlichen Kalender an der Geburt Christi orientiert, während die muslimische Zeitrechnung mit der Auswanderung Mohammeds von Mekka nach Medina beginnt. Künstliche Nullpunkte sind leicht dar-

an zu erkennen, dass das jeweilige Merkmal negative Ausprägungen besitzen kann, wie z. B. 1000 *vor* Christus.

4. **Ratioskala**: Das einfachste Erkennungszeichen ratioskalierter Merkmale ist die Existenz eines **natürlichen (echten) Nullpunktes**, der erst den **Vergleich von Verhältnissen** zwischen Skalenwerten ermöglicht. Das Alter (nicht Geburtsjahr!), das Einkommen und die Temperaturmessung in Kelvin (nicht Celsius oder Fahrenheit!) sind Eigenschaften auf Ratioskalenniveau. Hier können Verhältnisse interpretiert werden: Ein 50jähriger ist doppelt so alt wie ein 25jähriger. Die Temperaturmessung nach Kelvin ist im Gegensatz zu Celsius und Fahrenheit eine Ratioskala, da diese einen natürlichen Nullpunkt hat (bei -273,15 °C). Null Kelvin heißt: Abwesenheit von Temperatur bzw. Molekularbewegung, während 0 °Celsius eine definitorische Festlegung durch den Gefrierpunkt des Wassers ist. Bei 300 Kelvin ist es also tatsächlich doppelt so warm wie bei 150 Kelvin. Ratioskalierte Merkmale können keine negativen Werte annehmen. Es gibt weder ein negatives Einkommen noch ein negatives Alter und auch keine negative Temperatur in Kelvin.
   Ratioskalierte Merkmale besitzen „künstliche" Skaleneinheiten, was sich am Beispiel des Einkommens gut verdeutlichen lässt, das bis vor kurzem in DM und Pfennig, neuerdings aber in Euro und Cent gemessen wird. Dies unterscheidet Ratioskalen von Absolutskalen.
5. **Absolutskala**: Absolutskalen besitzen zusätzlich zu den bisher diskutierten Eigenschaften der anderen Skalen eine **natürliche Skaleneinheit**. Die Zuordnung der Zahlen ist durch die Beziehungen im empirischen Relativ eindeutig festgelegt. Absolutskalierte Merkmalsausprägungen besitzen z. B. alle Zählvariablen, wie die Semesterzahl, die Zahl der Bürgerkriege seit dem Zweiten Weltkrieg usw.

Bei nominal- oder ordinalskalierten Merkmalen handelt es sich immer um diskrete Merkmale, während metrische Merkmale diskret oder kontinuierlich sein können. Von einem *diskreten Merkmal* spricht man, wenn dieses abzählbar viele Werte annehmen kann. Als *kontinuierliche Merkmale* werden Merkmale bezeichnet, die in jedem beliebig kleinen Intervall unendlich (überabzählbar) viele Werte annehmen können. Die Semesterzahl ist beispielsweise ein diskretes, metrisches Merkmal; das Alter ein kontinuierliches, metrisches Merkmal. Kontinuierlich heißt zwar, dass das Merkmal unendlich viele Werte annehmen *kann*, nicht aber dass diese auch

gemessen werden können. So werden nur bestimmte Werte des Alters, z. B. ganze Jahre, gemessen; zwischen diesen gemessenen Werten existieren aber unendlich viele andere Werte. Statt von diskreten Merkmalen spricht man auch von *kategorialen* Merkmalen. Als *dichotom* werden Merkmale bezeichnet, die nur zwei Ausprägungen annehmen können, wie das Geschlecht. Merkmale mit mehr als zwei Ausprägungen werden häufig auch *polytom* genannt.

Die Einteilung in Skalenniveaus ist von besonderer Bedeutung für die statistische Auswertung der Daten. **Je höher das Skalenniveau, umso mehr statistische Verfahren sind zulässig**. So ist die Berechnung eines arithmetischen Mittels bei Nominalskalen und Ordinalskalen nicht erlaubt, da die Abstände zwischen den vergebenen Zahlen keine Bedeutung haben, die über die Unterschiedlichkeit bzw. die Ordnung hinausgeht. Von diesem Standpunkt ließe sich auch die Berechnung einer Durchschnittsnote kritisieren, da Schulnoten nur Ordinalskalenniveau aufweisen. Sie sagen ja nur etwas über einen Rang aus, nicht aber über die Abstände zwischen den Zahlen. Bei der Datenanalyse werden ordinalskalierte Merkmale allerdings häufig wie intervallskalierte Merkmale behandelt, um bestimmte statistische Verfahren anwenden zu können (vgl. dazu Allerbeck 1978). Man unterstellt dann, dass die Abstände der einzelnen Ausprägungen auf der Ordinalskala gleich sind. Multivariate Analyseverfahren (vereinfacht ausgedrückt: Verfahren, mit deren Hilfe gleichzeitig mehr als zwei Merkmale analysiert werden können) setzen häufig mindestens intervallskalierte Daten voraus – erwähnt seien hier stellvertretend die lineare Regressionsanalyse, die Faktorenanalyse und die Clusteranalyse (vgl. Backhaus et al. 2003). In den vergangenen Jahrzehnten wurden statistische Verfahren zur Analyse nominaler und ordinaler Daten (weiter)entwickelt und sind heute in den gängigen Statistik-Paketen implementiert (SPSS, Stata, SAS) (vgl. Andreß et al. 1997; Long 1997; Agresti 1996).

Welches Skalenniveau eine Variable annimmt, hängt neben den beobachtbaren Beziehungen zwischen den Objekten von der gewählten Operationalisierung ab. Misst man das Alter der in Mainz lehrenden Professorinnen und Professoren in Jahren, dann erhält man ein ratioskaliertes Merkmal. Genauso gut könnte man dem/der ältesten Professor/in die höchste Zahl aus einer beliebigen Reihe von Zahlen zuweisen, der/dem zweitältesten die zweithöchste usw. In diesem Fall hat man Alter auf Ordinalskalenniveau gemessen. Schließlich könnte man das Alter noch nominal messen, indem man nur zwischen Professorinnen und Professoren, die in der Zeit

des Nationalsozialismus geboren wurden, und anderen unterscheidet und den Ersteren z. B. eine 1, den Letzteren eine 2 zuweist.

Generell ist es sinnvoll, auf dem höchstmöglichen Skalenniveau zu messen, da höhere Skalenniveaus immer mehr Informationen enthalten als niedrigere. Hat man das Alter der Professorinnen und Professoren in Jahren gemessen, so kann man exakt angeben, um wie viele Jahre Professor/in X älter als Professor/in Y ist, während dies bei den anderen beiden Messungen nicht möglich ist. Zudem ist eine Verminderung des Skalenniveaus im Nachhinein immer möglich, nicht aber eine Erhöhung. Wählt man ein niedrigeres Skalenniveau als möglich, dann reduziert sich von Vorneherein die Zahl der zulässigen statistischen Verfahren. Gelegentlich kann es durchaus sinnvoll sein, auf einem niedrigeren Skalenniveau zu messen: Um die Antwortbereitschaft zu erhöhen, fragt man in Umfragen häufig nicht nach dem exakten Einkommen (Ratioskala), sondern gibt Kategorien vor (weniger als 500 Euro, 500 bis 1.000 Euro, ..., 5.000 Euro und mehr; Ordinalskala). Im Allbus wird das monatliche Nettoeinkommen zunächst offen abgefragt. Befragten, die die Auskunft verweigern, wird dann unter Zusicherung der Anonymität der Befragung eine Liste mit Einkommenskategorien vorgelegt (vgl. GESIS 2007, 274).

3.3 Skalierungsverfahren

Zur Messung komplexer Sachverhalte werden häufig mehrere Indikatoren herangezogen. So wird es kaum möglich sein, Konstrukte wie „Rechtsextremismus“ oder „Ausländerfeindlichkeit“ über einen einzigen Indikator angemessen zu erfassen. Die Verwendung mehrerer Indikatoren zur Messung einer interessierenden Dimension hat den Vorteil, dass die Messung zuverlässiger wird, wenn die Messfehler sich ausgleichen. Stellt sich nach einer Untersuchung heraus, dass ein Indikator den zu messenden Sachverhalt nicht gut abbildet, dann ist das bei Verwendung mehrerer Indikatoren kein so großes Problem. Vor allem ist bei Verwendung mehrerer Indikatoren deren Zuverlässigeit und Gültigkeit besser prüfbar (vgl. dazu Abschnitt 3.4).

Liegen mehrere Indikatoren vor, dann benötigen wir Verfahren zur Herstellung *eines* Messinstrumentes. Skalierungsverfahren sind nichts anderes als Verfahren zur Herstellung einer **Skala** aus mehreren Indikatoren. Auch mit einem **Index** können mehrere Indikatoren zusammengefasst werden.

Indizes und Skalen stellen in diesem Sinne „*zusammengesetzte Messungen*“ dar. Technisch gesprochen wird bei der Bildung eines Index/einer Skala aus mehreren Variablen eine neue Variable gebildet. Der Unterschied zwischen Skalen und Indizes besteht darin, dass bei Skalierungsverfahren die Dimensionalität der in die Skala eingehenden Indikatoren geprüft wird (vgl. Mayntz et al. 1978, 47). Indikatoren bilden nur dann eine Skala, wenn die Voraussetzungen des Skalierungsmodells (z. B. der Guttman-Skala) erfüllt sind.

Bei einem **Index** werden Indikatoren nach einer bestimmten mathematischen Anweisung zusammengefasst. Welche mathematische Operation zur Berechnung angewendet wird, hängt von der Fragestellung ab. Meistens werden additive Indizes verwendet, d. h. die Werte der einzelnen Variablen werden zur Bildung des Index einfach summiert. Indizes werden häufig verwendet, wenn mehrere Indikatoren, die unterschiedliche Dimensionen messen, zu einem neuen Instrument zusammengefasst werden. So könnte der *sozioökonomische Status* aus Indikatoren für die drei Dimensionen Bildung, Einkommen und Berufsprestige berechnet werden (vgl. Schnell et al. 2008, 167 ff.).

Bestandteile einer Skala (bzw. ganz generell Bestandteile eines Fragebogens) werden als **Items** bezeichnet. Dabei kann es sich um Statements oder Fragen handeln. Die verschiedenen Skalierungsverfahren unterscheiden sich vor allem danach, welche Anforderungen die Items erfüllen müssen, und wie diese zu einem einzigen Skalenwert verarbeitet werden. Außerdem können Skalierungsverfahren danach unterschieden werden, ob Personen und/oder Variablen skaliert werden.

An dieser Stelle wollen wir uns auf die Darstellung von zwei Skalen – der Likert- und der Guttman-Skala – beschränken. In älteren Studien findet man manchmal die *Thurstone-Skala*, die aber kaum Verwendung findet. Eine gute Darstellung dieser Skalierungsverfahren findet sich bei McIver und Carmines (1982).

3.3.1 Likert-Skala

Likert-Skalen werden in den Sozialwissenschaften zur Messung von Einstellungen eingesetzt. Beispiele für **Likert-Skalen** sind die Faschismus-, Antisemitismus- und Ethnozentrismus-Skala, die in der Untersuchung zum „Autoritären Charakter“ von Adorno, Frenkel-Brunswik, Levinson und

Sanford verwandt wurden (vgl. Adorno et al. 1950). Nehmen wir an, wir wollten eine Likert-Skala konstruieren, die „Ausländerfeindlichkeit“ messen soll. Zunächst müssen nun Indikatoren gefunden werden, die ausländerfeindliche Einstellungen messen. **Jeder einzelne Indikator soll dieselbe Dimension messen**, hier also negative Einstellungen zu Ausländern **und möglichst nichts anderes**.

Um die Konstruktion einer Likert-Skala zu erläutern, wird auf vier im ALLBUS 2006 enthaltene Aussagen zur Messung ausländerfeindlicher Vorurteile zurückgegriffen (Abbildung 3.2). Entwickelt man eine neue Skala, dann ist es sinnvoll, deutlich mehr Indikatoren als im Beispiel zu verwenden.

Abbildung 3.2: Messung ausländerfeindlicher Einstellungen

1. Die in Deutschland lebenden Ausländer sollten ihren Lebensstil ein bisschen besser an den der Deutschen anpassen.
2. Wenn Arbeitsplätze knapp werden, sollte man die in Deutschland lebenden Ausländer wieder in ihre Heimat zurückschicken.
3. Man sollte den in Deutschland lebenden Ausländern jede politische Betätigung in Deutschland untersagen.
4. Die in Deutschland lebenden Ausländer sollten sich ihre Ehepartner unter ihren eigenen Landsleuten auswählen.

Bei der Likert-Skalierung wird den Befragten die Möglichkeit eingeräumt, die Zustimmung bzw. Ablehnung der Statements in **abgestufter Form** vorzunehmen (*Rating-Format*). Im oben genannten Beispiel reicht das Spektrum über sieben Stufen, von „Stimme überhaupt nicht zu“ bis „Stimme voll und ganz zu“. Häufig werden bei einer Likert-Skala auch fünf Antwortalternativen vorgegeben: lehne stark ab, lehne ab, teils/teils, stimme zu, stimme stark zu.

Die **Zahlenzuordnung zu den Antwortalternativen** erfolgt am sinnvollsten so, dass eine starke Ausprägung auf der zu messenden Einstellung einen hohen Zahlenwert erhält und umgekehrt. Ausländerfeindlichen Einstellungen werden also hohe Zahlenwerte zugeordnet, bei nichtausländerfeindlichen Einstellungen niedrige. Die Antwort „stimme voll und ganz zu“ gibt bei den Beispielitems immer eine ausländerfeindliche Einstellung wieder, d. h. die Fragen sind alle in die gleiche Einstellungsrichtung „gepolt“ (*gleichsinnig*). Der Antwort „stimme voll und ganz zu“ wird

dementsprechend bei allen Aussagen die Zahl 7, der Antwort „stimme überhaupt nicht zu“ die Zahl 1 und den dazwischen liegenden Abstufungen die Zahlen 2 bis 6 zugeordnet.

Die Beurteilung der Aussagen durch die Befragten findet sich in Abbildung 3.3. Der Aussage, Ausländer sollten ihren Lebensstil etwas besser an den der Deutschen anpassen, wurde am stärksten zugestimmt. Knapp 50 % der Befragten stimmten der Aussage voll und ganz zu (Skalenwert 7). Am stärksten abgelehnt wurde die Aussage, Ausländer sollten Ihre Ehepartner unter ihren Landsleuten auswählen. Mehr als 50 % der Befragten lehnten die Aussage voll und ganz ab. Beide Items sind schief verteilt, die Aussage zur Lebensstilanpassung linksschief, die Aussage zu Ehepartnern rechtsschief.

Abbildung 3.3: Einstellungen gegenüber Ausländern

Skala von 1, stimme überhaupt nicht zu, bis 7, stimme voll und ganz zu
Daten: ALLBUS 2006, Westdeutsche, N=2.038

Zustimmende Äußerungen geben hier bei allen vier Items negative Einstellungen gegenüber Ausländern wider. Es kann jedoch sinnvoll sein, positive

und negative Statements zu formulieren, um Zustimmungstendenzen kontrollieren zu können. Unter Zustimmungstendenzen wird die Neigung von Befragten verstanden, Aussagen ohne Berücksichtigung ihres Inhalts zuzustimmen (vgl. auch Kapitel 4.1). Zustimmungstendenzen lassen sich daran erkennen, dass ein Befragter positiv wie negativ formulierten Aussagen zustimmt, was inhaltlich nicht plausibel ist. Ein positives Statement zu Ausländern wäre ein zusätzlichen Item wie „Die politischen Einflussmöglichkeiten der in Deutschland lebenden Ausländer sollten gestärkt werden". Zustimmung bedeutet hier gerade die Abwesenheit von negativen Einstellungen zu Ausländern. Will man positiv und negativ formulierte Aussagen zu einer Skala zusammenfassen, muss man darauf achten, dass dieselbe Zahl dieselbe Einstellung repräsentiert. Z. B. indem man bei positiv formulierten Aussagen der Antwort „Stimme voll und ganz zu" den Wert 1 (Abwesenheit von Ausländerfeindlichkeit) und der Antwort „Stimme überhaupt nicht zu" den Wert 7 (Ausländerfeindlichkeit) zuweist und die Abstufungen entsprechend rekodiert (*umpolt*).

Die Konstruktion einer Likert-Skala lässt sich anhand der im ALLBUS 2006 verwendeten Items veranschaulichen. Zur Vereinfachung werden ausschließlich die 2.038 in Westdeutschland befragten Personen betrachtet, die alle vier Aussagen auf der siebenstufigen Skala beantwortet haben, also keinen einzigen fehlenden Wert (keine Angabe/weiß nicht) aufweisen.[1] Bei der Likert-Skala wird der Skalenwert aus der Summe aller (gleich gepolten) Items berechnet. Die Addition ist jedoch nur dann gerechtfertigt, wenn die Items eine einzige Dimension messen. Anhand der **Item-Analyse** wird entschieden, welche Items geeignet sind und damit in die endgültige Skala eingehen.

Der Item-Analyse liegen zwei Gedanken zugrunde. Messen die Items dieselbe Dimension, dann sollten sich Unterschiede auf der zu messenden Dimension auch in unterschiedlichen Antworten niederschlagen. Zudem sollten die Antworten der Befragten zu den einzelnen Statements dann konsistent sein.

1 Dieses Vorgehen wird auch als listenweiser Fallausschluss (*listwise deletion*) bezeichnet. Der listenweise Fallausschluss wird in empirischen Analysen häufig praktiziert, führt aber nur unter bestimmten Bedingungen zu unverzerrten Ergebnissen. Nämlich dann, wenn die kompletten Fälle als eine Zufallsstichprobe aus allen Fällen betrachtet werden können (*missing completely at random*, vgl. einführend Schafer und Graham 2002).

Um zu überprüfen, ob sich Unterschiede in der zu messenden Einstellung auch in der Beantwortung der einzelnen Aussagen niederschlagen, teilt man die Befragten in **Extremgruppen** auf. Zunächst berechnet man für jede Person die Summe der Werte über alle Items. Im Beispiel ist der niedrigste mögliche Wert der Summe 4 (wenn bei allen vier Items der Wert 1 vorliegt), der höchste mögliche Wert beträgt 28 (wenn bei allen vier Items der Wert 7 vorliegt). Man wählt dann diejenigen 25 % der Befragten aus, die die niedrigsten Werte über alle Items aufweisen und diejenigen 25 % der Befragten mit den höchsten Werten. Für die vier ALLBUS-Items haben die 25 % der Befragten mit den niedrigsten Werten Werte zwischen 4 und 11 (Gruppe 1). Die 25 % der Befragten mit den höchsten Werten haben Werte zwischen 18 und 28 (Gruppe 2).

Danach vergleicht man die Antworten der beiden Extremgruppen *zu jedem einzelnen Item*. Brauchbar sind diejenigen Items, bei denen sich die Antworten der Extremgruppen unterscheiden. Tabelle 3.1 beinhaltet die Durchschnittswerte der beiden Extremgruppen ($\bar{x}_1$ und $\bar{x}_2$) für die einzelnen Items. Befragte mit extrem hohen Werten auf allen Items (Gruppe 2) sollten auch jedem einzelnen Item deutlich stärker zustimmen als Befragte mit extrem niedrigen Werten auf allen Items (Gruppe 1). Dies ist auch der Fall. Am stärksten unterscheiden sich die beiden Gruppen in der Beantwortung der Frage, ob man Ausländern jede politische Betätigung in Deutschland untersagen sollte. Am geringsten ist der Unterschied in der Beurteilung der Aussage, Ausländer sollten ihren Lebensstil besser an den der Deutschen anpassen.

Bei der Analyse beschränkt man sich nicht auf einen Vergleich der Mittelwerte. Vielmehr wird für jedes Item ein Trennschärfe-Index berechnet, der dem t-test für Mittelwertunterschiede entspricht (Kapitel 12.3.1). Trennschärfe-Indizes größer als 1,65 gelten als ausreichend zur Annahme eines Unterschiedes in der Beantwortung der Items durch die beiden Extremgruppen. Wie man sieht, sind die Trennschärfe-Indizes alle deutlich größer als 1,65. Die Items scheinen daher ein und dieselbe Dimension zu messen und zur Konstruktion einer Skala geeignet.

Eine andere Methode der Itemanalyse ist die Berechnung von **Trennschärfe-Koeffizienten**. Messen alle Aussagen ein und denselben Sachverhalt, dann sollten die Items hoch miteinander korrelieren. Ein Befragter, der der Aussage „Die in Deutschland lebenden Ausländer sollten ihren Lebensstil ein bisschen besser an den der Deutschen anpassen“ stark

Tabelle 3.1: Extremgruppenanalyse

	Gruppe 1	Gruppe 2	Differenz	
	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_2 - \bar{x}_1$	t-Wert
Lebensstilanpassung	4,5	6,6	2,1	25
	(s_1^2=1,723)	(s_2^2=0,884)		
Arbeitsplätze	1,3	5,1	3,8	51
	(s_1^2=0,636)	(s_2^2=1,614)		
Politische Betätigung	1,3	5,5	4,2	57
	(s_1^2=0,706)	(s_2^2=1,600)		
Ehepartner	1,1	4,2	3,1	35
	(s_1^2=0,478)	(s_2^2=2,01)		
Anzahl der Befragten	553	557		

Daten: Allbus 2006, Westdeutsche. $\bar{x}$ = arithmetisches Mittel, s^2=Varianz

zustimmt, müsste auch das Statement „Man sollte den in Deutschland lebenden Ausländern jede politische Betätigung in Deutschland untersagen" befürworten. Unplausibel wäre dagegen die Zustimmung zum ersten Statement bei gleichzeitiger Ablehnung des zweiten Statements. Treten solche inkonsistenten Antwortmuster häufiger auf, muss man vermuten, dass die Statements Unterschiedliches messen. Bei inhaltlich plausiblen Antworten ist der Zusammenhang zwischen der Beantwortung eines Items und allen anderen Items dagegen sehr stark. Um dies zu überprüfen, berechnet man die Stärke des Zusammenhangs (vgl. Kapitel 7.6, Pearsons r) zwischen jedem einzelnen Item und dem Summenwert über alle Items (Spalte 4). Dieses Maß wird als Trennschärfe-Koeffizient bezeichnet. Der Trennschärfe-Koeffizient kann maximal eins werden.[2] Nämlich dann, wenn es einen perfekten positiven Zusammenhang zwischen Item und Skala gibt. Statements, die nur schwach mit den anderen Statements zusammenhängen, sind für die Skala untauglich und werden in der endgültigen Skala nicht verwendet. Die Korrelationskoeffzienten der Items mit der Skala bewegen sich zwischen 0,55 und 0,79. Angegeben wird zudem der korrigierte Trenschärfe-Koeffizient (Spalte 5). Dieser gibt die Korrelation eines Items mit der Summe aller restlichen Items an. Diese Korrektur wird vorgenommen, weil die Korrelation zwischen Item und Skala dadurch, dass das Item auch in der Skala enthalten ist, künstlich überschätzt wird. Man

2 Bei gleichsinnig gepolten Items kann der Koeffizient nur positive Werte annehmen.

sieht, dass die korrigierten Trennschärfe-Koeffizienten deutlich geringer sind. Die Aussage, Ausländer sollten ihren Lebensstil besser an den der Deutschen anpassen, ist weniger gut als die anderen Aussagen geeignet, um Personen mit und ohne negative Einstellungen gegenüber Ausländern zu unterscheiden. In der letzten Spalte ist Cronbachs α angegeben, auf dessen Berechnung und Interpretation weiter unten eingegangen wird.

Tabelle 3.2: Trennschärfe-Koeffizienten und Cronbachs α

			Korrelation mit		Cronbachs α
Item	$\bar{x}$	s	Skala	Rest-Skala	(ohne Item)
Lebensstilanpassung	5,7	1,55	0,55	0,30	(0,70)
Arbeitsplätze	3,1	1,92	0,78	0,56	(0,54)
Politische Betätigung	3,3	2,12	0,79	0,53	(0,56)
Ehepartner	2,3	1,86	0,71	0,45	(0,62)
Skala	14,4	5,34			0,68

$\bar{x}$= arithmetisches Mittel, s=Standardabweichung
Daten: Allbus 2006, westdeutsche Befragte, n=2.038.

Beide Verfahren können zu unterschiedlichen Ergebnissen führen, da in einem Fall „nur“ die Extremgruppen, im anderen Fall alle Befragten in die Berechnung einfließen. Die Methode der Extremgruppen ist das ältere, die Berechnung von Trennschärfe-Koeffizienten das neuere Verfahren.

Nachdem die Items für die endgültige Skala ausgewählt sind, kann der **Skalenwert für jede Person berechnet** werden. Dies geschieht durch Addition der Werte der ausgewählten Items. Werden trotz des niedrigen Trennschärfe-Koeffizienten für das Item Lebensstilanpassung alle Items zur Bildung der Skala herangezogen, so ist der niedrigste mögliche Wert 4, der inhaltlich der Abwesenheit von negativen Einstellungen zu Ausländern entspricht, und der höchste Wert 28, der die höchstmögliche Ausprägung negativer Einstellungen zu Ausländern wiedergibt. Damit ist die Likert-Skala konstruiert und das Messinstrument ist fertig. Der Mittelwert auf der Skala beträgt 14, die Standardabweichung – ein Maß für die Streuung der Antworten (Kapitel 6.2.4) – 5,34.

So einfach die Berechnungsanweisung ist, so schwierig ist die **Interpretation** der Skalenwerte – zumindest im mittleren Bereich der Skala. Die

Position eines Befragten muss außerdem relativ zu allen anderen interpretiert werden. Bei einem durchschnittlichen Skalenwert von 14 ist 10 ein niedriger Skalenwert. Bei einem Durchschnittswert von 9 ist ein Skalenwert von 10 dagegen relativ „normal". Zudem muss berücksichtigt werden, ob das Antwortverhalten einer Gruppe eher homogen oder heterogen ist.[3]

Auf einen Punkt muss hingewiesen werden: Um die Itemanalyse durchführen zu können, gehen wir davon aus, dass die Abstände zwischen den einzelnen Skalenpunkten gleich groß sind, und damit z. B. die Extremantworten („Stimme überhaupt nicht zu", „Stimme voll und ganz zu") gleich weit von der Mitte entfernt sind. Wir nehmen also an, dass die zur Itemanalyse herangezogenen Items mindestens intervallskaliert sind. Ohne diese Annahme dürften wir arithmetisches Mittel, Varianz, t-test (Trennschärfe-Index) und Pearsons r (Trennschärfe-Koeffizient) nicht berechnen. Ob die Abstände zwischen den Kategorien äquidistant sind, kann bei mehreren Items mit Ratingskalen-Format geprüft werden (vgl. Rost 2004, Kapitel 3).

In der empirischen Sozialforschung werden häufig eine Reihe von Items als Likert-Skalen bezeichnet, weil sie fünfstufige Antwortalternativen haben. Auch Summenindizes werden manchmal Likert-Skalen genannt. Um Likert-Skalen handelt es sich jedoch nur dann, wenn vor Addition der Werte geprüft wurde, ob die Skala eindimensional ist. Faktorenanalysen sind eine alternative Methode zur Prüfung der Eindimensionalität.

3.3.2 Guttman-Skala

Die **Guttman-Skala** unterscheidet sich in der Konstruktion deutlich von der Likert-Skala. Durch die Guttman-Skalierung werden gleichzeitig Personen und Aussagen hinsichtlich der zu messenden Dimension in eine Rangfolge gebracht werden. Die Skalen für konventionelle und unkonventionelle Partizipation der *Political Action Studie* (vgl. Barnes et al. 1979) sind Beispiele für Guttman-Skalen.

Einer Guttman-Skala liegen **Items** zugrunde, die **hinsichtlich der zu messenden Dimension immer extremer** werden, d. h. die Dimension

3 Man kann die Skalenwerte standardisieren, indem man eine z-Transformation (vgl. Gleichung 10.13, S. 242) durchführt. Die z-Werte geben die Abweichung des Skalenwertes eines Befragten vom durchschnittlichen Skalenwert in Abhängigkeit von der Streuung der Skalenwerte wieder.

in einer unterschiedlichen Intensität messen. Als Beispiel werden Fragen zur politischen Partizipation aus dem ALLBUS 1998 herangezogen. Unter anderem wurde die *Bereitschaft zu unkonventioneller politischer Partizipation* gemessen. Die Frage lautete: „Wenn Sie politisch in einer Sache, die Ihnen wichtig ist, Einfluß nehmen, Ihren Standpunkt zur Geltung bringen wollten: Welche der Möglichkeiten auf diesen Karten würden Sie dann nutzen, was davon käme für Sie in Frage?“ Auf den Karten waren eine Reihe konventioneller (Wählen, Mitarbeit in einer Partei usw.) und unkonventioneller Partizipationsformen angegeben (vgl. Zentralarchiv für empirische Sozialforschung 1999, 60–68). Aus den vorgegebenen Items haben wir für das Beispiel drei Indikatoren ausgewählt (Abbildung 3.4)

Abbildung 3.4: Messung unkonventioneller politischer Partizipation

1.	Teilnahme an einer nicht genehmigten Demonstration
2.	Mitarbeit in einer Bürgerinitiative
3.	Hausbesetzung, Besetzung von Fabriken, Ämtern

Schwierig ist ein Item dem wenige Befragte zustimmen. Ein Item ist leicht, wenn es von vielen Befragten bejaht wird. Nach dem Schwierigkeitsgrad werden die Items in eine Reihenfolge gebracht. Zur Erläuterung der Logik des Verfahrens nehmen wir vorerst an, dass die Items von Bürgerinitiativen über ungenehmigte Demonstrationen zu Hausbesetzungen immer schwieriger werden. Durch die Befürwortung von Hausbesetzungen wird also eine größere Bereitschaft zu unkonventioneller Beteiligung ausgedrückt als durch die Befürwortung ungenehmigter Demonstrationen, und durch die Befürwortung ungenehmigter Demonstrationen eine größere als durch Bürgerinitiativen.

Die **Antwortvorgaben** bei den Items einer Guttman-Skala sind **dichotom**, d. h. es werden nur zwei Antwortmöglichkeiten – Zustimmung oder Ablehnung – vorgegeben. Angenommen wird, dass bis zu einem gewissen Schwellenwert der zu messenden Eigenschaft (hier: Befürwortung unkonventioneller politischer Partizipation) ein Item abgelehnt wird. Überschreitet die zu messende Einstellung diesen Schwellenwert, dann wird das Item befürwortet. Bis zu einem bestimmten Ausmaß der Befürwortung unkonventioneller politischer Partizipation wird z. B. die Beteiligung an einer Bürgerinitiative abgelehnt. Erst wenn unkonventionelle Beteiligungsformen in einem bestimmten Ausmaß befürwortet werden, wird die

Beteiligung an einer Bürgerinitiative bejaht. Bis zu einem bestimmten Schwellenwert der zu messenden Eigenschaft ist die Wahrscheinlichkeit der Befürwortung eines Items null, ab diesem Schwellenwert ist die Wahrscheinlichkeit eins. Die Wahrscheinlichkeit der Befürwortung steigt also bei einer bestimmten Ausprägung der zu messenden Eigenschaft sprunghaft von null auf eins an.

Messen die Items eine einzige Dimension in unterschiedlicher Intensität, dann kann angenommen werden, dass eine Person, die ein bestimmtes Item bejaht, alle weniger extremen Items ebenfalls bejaht. Stimmt die oben angenommene Reihenfolge, dann ist anzunehmen, dass ein Befragter, der an Hausbesetzungen teilnehmen würde, ebenfalls die Teilnahme an ungenehmigten Demonstrationen und an Bürgerinitiativen in Betracht zieht. Andererseits dürfte eine Person, die sich nicht an Bürgerinitiativen beteiligen möchte, keine ungenehmigten Demonstrationen und erst recht keine Hausbesetzungen erwägen. Die Items sollten also eine kumulative Ordnung aufweisen. Bei einer idealen Guttman-Skala kann aus der Summe der bejahten Items nicht nur geschlossen werden, wie vielen, sondern auch welchen Items der Befragte zugestimmt hat: bei einem bejahten Item dem schwächsten Item, bei zwei bejahten Items den beiden schwächsten Items usw. Ideal soll heißen, dass nur modellkonforme Antwortmuster auftreten. Aus der Anzahl bejahter Items kann dann das Antwortmuster genau reproduziert werden.

Bei drei Items – wie in unserem Beispiel – existieren vier *zulässige* bzw. *modellkonforme* Antwortmuster. Diese sind in Tabelle 3.3 auf der folgenden Seite dargestellt. + bedeutet die Zustimmung zu einer Aussage, − deren Ablehnung; Item 1 kennzeichnet das schwächste, Item 2 das mittlere und Item 3 das extremste Item. Das erste Antwortmuster (− − −) kennzeichnet die Ablehnung aller Items. Das zweite Antwortmuster (+ − −) gibt an, dass dem schwächsten Item (Item 1) zugestimmt, die beiden extremeren Items (Item 2 und Item 3) dagegen ablehnt wurden. Der Skalenwert entspricht der Zahl des „extremsten“ Items, das bejaht wurde. Im dritten Antwortmuster ist dies das zweite Item; für dieses Antwortmuster wird also der Skalenwert „2“ vergeben.

Ob die Items nun eine Guttman-Skala bilden, lässt sich daran ablesen, wie gut die Antwortmuster aus der Anzahl bejahter Items reproduziert werden können. Die in Tabelle 3.4 auf der nächsten Seite dargestellten Antwortmuster entsprechen nicht den Anforderungen der Guttman-Skala. Bei die-

Tabelle 3.3: Modellkonforme Antwortmuster bei der Guttman-Skala

Item 1	Item 2	Item 3	Skalenwert
−	−	−	0
+	−	−	1
+	+	−	2
+	+	+	3

sen Antwortmustern lässt sich aus der Anzahl bejahter Items nicht mehr ablesen, welche Items bejaht wurden. Bei nur einem bejahten Item muss man nach der Logik der Guttman-Skala eigentlich davon ausgehen, dass Item 1 (das am wenigsten extreme Item) bejaht und die beiden extremeren Items verneint wurden (Antwortmuster: $+--$). Tatsächlich wurde im ersten Antwortmuster jedoch nur das zweite Item bejaht (Antwortmuster: $-+-$). Wenn man aus der Anzahl bejahter Items das Antwortmuster ableitet, begeht man also zwei Fehler: Die Beantwortung des ersten Items ($-$ statt $+$) und des zweiten Items ($+$ statt $-$) werden falsch eingeschätzt.[4] Messen die Items eine Dimension in unterschiedlicher Intensität, dann kann es eigentlich nicht sein, dass dem extremsten Item zugestimmt wird, die beiden weniger extremen Items jedoch abgelehnt werden ($- - +$).

Tabelle 3.4: Nicht modellkonforme Antwortmuster bei der Guttman-Skala

Item 1	Item 2	Item 3	Wert	Fehler
−	+	−	1	2
−	−	+	1	2
−	+	+	2	2
+	−	+	2	2

Wert = Anzahl bejahter Items

Mit der **Skalogramm-Analyse** wird anhand der Antwortmuster geprüft, ob die Items zur Bildung einer Guttman-Skala geeignet sind. Je höher der Anteil unzulässiger Antwortmuster, umso ungeeigneter sind die Items.

4 Zur Berechnung der Fehler existieren mehrere Verfahren. Hier werden die Fehler als Abweichung vom idealen Muster berechnet (Methode von Goodenough und Edwards, vgl. McIver und Carmines 1982, 42).

Das Verfahren soll an den ausgewählten Items zur politischen Beteiligung demonstriert werden.

Zunächst müssen wir feststellen, welche Items extremer und welche weniger extrem sind. Diese Entscheidung fällt nicht aufgrund vorheriger Überlegungen (auch wenn man sich schon bei der Formulierung der Fragen Gedanken über deren Intensität macht), sondern aufgrund der Antworten der Befragten. **Die Reihenfolge der Statements ist also eine empirische Frage**. Die Items werden also zunächst nach der **Häufigkeit der Zustimmungen geordnet**. Von den 2.117 westdeutschen Befragten im ALLBUS 1998 gaben ca. 30 % (646) an, dass sie gegebenenfalls an einer Bürgerinitiative mitarbeiten würden, knapp 10 % (202) zogen die Beteiligung an einer ungenehmigten Demonstration in Betracht, während lediglich 3 % (76) der Befragten zur Besetzung von Häusern bereit wären. Die Reihenfolge der Items ist also: Bürgerinitiative, ungenehmigte Demonstration, Hausbesetzung.

Nun muss ermittelt werden, wie viele der Antworten nicht dem idealen Muster entsprechen. In der ersten Spalte von Tabelle 3.5 sind im oberen Teil die modellkonformen und, im unteren Teil, die nicht-modellkonformen Antwortmuster angegeben. Im Kopf der ersten Spalte finden sich die drei Items, geordnet nach deren Schwierigkeit, die durch die prozentuale Zustimmung ermittelt wurde. In der zweiten Spalte ist die Zahl der bejahten Items, in der dritten die Zahl der Fehler pro Antwortmuster wiedergegeben. Die Frage ist nun, wie viele der Befragten modellkonform antworteten und wie viele nicht. Wie man in der vierten Spalte sehen kann, antworteten 2.025 Befragte (1.389+505+90+41) modellkonform; 92 (57+11+14+10) Befragte antworteten nicht im Sinne des Modells. Insgesamt (vgl. die letzte Spalte der Tabelle) werden bei der Vorhersage des Antwortmusters aus der Zahl bejahter Items 184 Fehler gemacht.

Bei einer idealen Guttman-Skala würden keine Fehler auftreten und unsere Vorhersagegenauigkeit betrüge 100 %. Der Anteil modellkonformer Antworten (= die zulässigen Antwortmuster) an allen Antworten wird als *Reproduzierbarkeitskoeffizient* bezeichnet.

$$\text{Rep.} = 1 - \frac{\text{Anzahl der Fehler}}{\text{alle Antworten}}$$

Tabelle 3.5: Guttman Skala – Politische Beteiligung 1998

BI	UD	HB	Wert	Fehler	Befragte	Summe der Fehler
Zustimmung						
30 %	10 %	3 %				
−	−	−	0	0	1389	
+	−	−	1	0	505	
+	+	−	2	0	90	
+	+	+	3	0	41	
−	+	−	1	2	57	57×2
−	−	+	1	2	11	$+ 11 \times 2$
−	+	+	2	2	14	$+ 14 \times 2$
+	−	+	2	2	10	$+ 10 \times 2$
						$= 184$

BI: Bürgerinitiative; UD: ungenehmigte Demonstration; HB: Hausbesetzung, Besetzung von Fabriken und Ämtern
Quelle: ALLBUS 1998, westdeutsche Befragte

$$\text{Rep.} = 1 - \frac{\text{Anzahl der Fehler}}{\text{Anzahl der Befragten} \times \text{Anzahl der Items}}$$

Der Reproduzierbarkeitskoeffizient ist also ein Maß für die Güte der Skala. Ein objektives Kriterium für die notwendige Größe dieses Koeffizienten existiert nicht. Als Faustregel hat sich ein Wert von 0,9 – also eine Vorhersagegenauigkeit von mindestens 90 % – eingebürgert.

In unserem Beispiel ergibt sich:

$$\text{Rep.} = 1 - \frac{184}{2117 \times 3} = 0,97$$

Oder anders ausgedrückt: Im Beispiel lassen sich 97 % der Antworten aufgrund der Skalenwerte richtig vorhersagen. Die drei verwendeten Items sind nach diesem Maß zur Bildung einer Guttman-Skala geeignet.

Sind die Items geeignet, dann wird für jeden Befragten der **Skalenwert** berechnet. Beim hier gewählten Verfahren (McIver und Carmines 1982,

51 f.) ergibt sich der Skalenwert aus der Anzahl bejahter Items. Wir erhalten eine stark schiefe Verteilung der Skalenwerte, da ein erheblicher Prozentsatz der Befragten überhaupt keine der vorgegebenen unkonventionellen Beteiligungsformen in Betracht zieht. Die Guttman-Skala misst auf **ordinalem Niveau**, da keine Annahme über die Skalenabstände gemacht werden. Zur Vereinfachung der Darstellung wurden lediglich drei Items betrachtet. In wissenschaftlichen Analysen sollten – wenn möglich – mehr Indikatoren herangezogen werden.

Beachtet werden muss, dass die Skalierbarkeit eine empirische Frage ist und damit von der jeweiligen Datenbasis abhängt. Es ist möglich, dass dieselben Items in einer Umfrage eine Likert-/Guttman-Skala bilden, während sie in einer anderen Umfrage die Skalenvoraussetzungen nicht erfüllen. Bei international oder interkulturell vergleichenden Analysen Harkness et al. (2003) sollte die Skalierbarkeit für jede Gruppe getrennt untersucht werden.

3.4 Gütekriterien einer Messung

Als Anhaltspunkte für die Qualität einer Messung werden deren Reliabilität und Validität herangezogen (vgl. Carmines und Zeller 1979). Unter **Reliabilität** wird die *Zuverlässigkeit* einer Messung verstanden. Ein Messinstrument ist umso zuverlässiger, je kleiner der Messfehler ist. Mit **Validität** bezeichnet man die *Gültigkeit* einer Messung. Hier geht es darum, ob tatsächlich das gemessen wurde, was gemessen werden sollte, und nicht irgendetwas anderes.

3.4.1 Reliabilität

Was Reliabilität eines Messinstrumentes bedeutet, kann man sich an folgendem Beispiel deutlich machen: Nehmen wir an, wir wollen die Temperatur messen. Wenn wir zwei Messinstrumente haben, die die Temperatur messen – z. B. ein Thermometer und ein Bimetall, das sich bei Wärme ausdehnt und bei Kälte zusammenzieht –, sollten Messungen mit diesen Instrumenten zu demselben Ergebnis kommen. Mit dem Thermometer messen wir eine Temperatur von 24 °C. Eine parallel durchgeführte Messung mit dem Bimetall ergibt eine Temperatur von 21 °C. Wenn sichergestellt ist, dass die unterschiedlichen Messergebnisse nicht andere Gründe haben (Sonneneinstrahlung, Wind usw.), dann muss man davon ausgehen,

dass eine der Messungen bzw. eines der Messinstrumente nicht besonders zuverlässig die Temperatur misst, eventuell sogar beide. *Die Messinstrumente weisen keine hohe Reliabilität auf.*

Wir könnten auch nur mit einem Instrument, z. B. einem Thermometer, arbeiten, und kurz nach unserer ersten Messung eine zweite durchführen. Wenn die Temperatur sich zwischen den beiden Zeitpunkten nicht geändert hat, z. B. weil der zeitliche Abstand zwischen den Messungen sehr gering ist, das Thermometer aber trotzdem unterschiedliche Werte anzeigt, kann man auch hier davon ausgehen, dass das Instrument nicht reliabel ist. Mit diesem Thermometer erhält man keine zuverlässigen Messwerte. Beide Überprüfungsmethoden werden auch im sozialwissenschaftlichen Bereich angewendet, zudem kann die interne Konsistenz eines Messinstrumentes geprüft werden (vgl. auch zu neueren Methoden Rost 2004, 376-380):

- Paralleltestverfahren
- Test-Retest-Verfahren
- Interne Konsistenz

Beim **Paralleltestverfahren** wird die Zuverlässigkeit durch zwei verschiedene Messinstrumente geprüft, die dasselbe messen sollen. Je stärker beide Messungen miteinander korrelieren (vgl. Kapitel 7), umso höher ist die Reliabilität der Messungen. Weichen die Ergebnisse stark voneinander ab, dann sind die Messinstrumente nicht reliabel. Allerdings existieren nur selten vergleichbare (= parallele) Messinstrumente. Ein Thermometer kann andere Messergebnisse produzieren als ein Bimetall, zumindest genauere. Bei sozialwissenschaftlichen Fragestellungen stellt sich dieses Problem noch viel gravierender. So kann zum Beispiel ein Messinstrument für „Nationalstolz“ die Frage sein: „Sind Sie stolz, Deutscher zu sein?“. Ein anderes Instrument könnte die Frage sein: „Sind Sie stolz auf Deutschland?“. Die unterschiedlichen Frageformulierungen könnten der Grund für unterschiedliche Ergebnisse sein. Man kann anzweifeln, on die beiden Fragen wirklich dasselbe messen, selbst wenn man davon ausgeht, dass beide Fragen „Nationalstolz“ erfassen.

Mit dem **Test-Retest-Verfahren** wird die Reliabilität durch die wiederholte Anwendung des Messinstruments geprüft. Führen zwei nacheinander erfolgte Messungen mit demselben Instrument zu unterschiedlichen Ergebnissen, dann misst das Messinstrument nicht zuverlässig. Allerdings muss

man von der *Stabilität des wahren Wertes* (im obigen Beispiel die Temperatur) zwischen den zwei Messzeitpunkten ausgehen, um unterschiedliche Messergebnisse auf die mangelnde Reliabilität des Instrumentes zurückführen zu können. Auf das Beispiel der Messung von „Nationalstolz" angewendet, hieße dies, dass die Frage „Sind Sie stolz, Deutscher zu sein?" reliabel misst, wenn die Antworten zu unterschiedlichen Zeitpunkten bei denselben Befragten hoch miteinander korrelieren. Vorausgesetzt, der Nationalstolz hat sich zwischen beiden Zeitpunkten nicht verändert. Um dies zu gewährleisten, kann man die beiden Messungen möglichst zeitnah durchführen. Erinnern sich die Befragten an ihre erste Antwort und versuchen sie, möglichst übereinstimmend zu antworten, dann wird die Reliabilität des Instruments allerdings überschätzt.

Am besten lässt sich die Reliabilität überprüfen, wenn ein Messinstrument aus mehreren Indikatoren besteht, die alle dieselbe Dimension messen sollen – so wie es beispielsweise bei den oben genannten vier Indikatoren zur Messung ausländerfeindlicher Einstellungen aus dem ALLBUS 2006 der Fall ist. Besteht das Instrument aus mehreren Indikatoren, dann kann die **interne Konsistenz** der Einzelmessungen überprüft werden. Zur Überprüfung der internen Konsistenz wird meistens Cronbachs α verwendet, dessen Berechnung aber ebenfalls intervallskalierte Indikatoren voraussetzt. Der Wert von α ist abhängig von der mittleren Interkorrelation der Items und der Zahl der Items. Je mehr Items in die Berechnung einfließen, umso höher ist Cronbachs α bei gleicher mittlerer Interkorrelation. Cronbachs α sollte größer als 0,8 sein, in empirischen Analysen werden aber auch geringere Werte akzeptiert (vgl. die Beispiele bei Diekmann 2008, 254).

Cronbachs α beträgt für die vier ALLBUS-Items 0,68 (Tabelle 3.2).[5] In den Zeilen der Einzelitems finden sich zudem noch die Werte von Cronbachs α, wenn das jeweilige Item nicht in die Berechnung der Skala einginge. Man sieht, dass die Reliabilität der Skala ohne das Item Lebensstilanpassung

5 Die Formel zur Berechnung lautet: $\alpha = \frac{p}{p-1}\left(1 - \frac{\sum_{i=1}^{p} s_i^2}{s_{\text{Skala}}^2}\right)$. p ist die Zahl der Items. Im Zähler des zweiten Bruchs steht die Summe der Varianzen der Einzelitems, im Nenner die Varianz der Skala. Je stärker die Items zusammen hängen, umso größer wird s_{Skala}^2 im Vergleich zu $\sum s_i^2$. Ist der Zusammenhang zwischen den Items null, dann entspricht die Varianz der Skala der Summe der Varianzen der Einzelitems und Cronbachs α nimmt den Wert null an. Für das Beispiel ergibt sich $\alpha = \frac{4}{4-1}\left(1 - \frac{1{,}55^2+1{,}92^2+2{,}12^2+1{,}86^2}{5{,}34^2}\right) = 0,68$.

etwas höher ausfallen würde, jedoch deutlich schlechter wäre, wenn die Indikatoren Arbeitsplätze oder politische Betätigung nicht in der Skala enthalten wären.

Festzuhalten bleibt, dass die Überprüfung der Reliabilität nicht einfach ist. Beim Paralleltestverfahren müssen vergleichbare Messinstrumente verwendet werden, beim Test-Retest-Verfahren muss sichergestellt sein, dass der wahre Wert unverändert geblieben ist. Zur Überprüfung der internen Konsistenz benötigt man keine der beiden Annahmen. Dies ist einer der Hauptvorteile, die die Messung einer Dimension durch mehrere Indikatoren mit sich bringt.

3.4.2 Validität

Während sich Reliabilität auf den technischen Aspekt einer Messung bezieht, betrifft die Validität den inhaltlichen Aspekt.

Am Beispiel der Parteiidentifikation soll die Validität einer Messung verdeutlicht werden (vgl. bereits Falter 1977). Unter der Parteiidentifikation wird in der Wahlforschung eine langfristig stabile, psychologische Bindung an eine Partei verstanden (vgl. Campbell et al. 1980, 121). Die Parteiidentifikation wird in der Bundesrepublik durch folgende Frage gemessen: „Viele Leute neigen in der Bundesrepublik längere Zeit einer bestimmten Partei zu, obwohl sie auch ab und zu eine andere Partei wählen. Wie ist das bei Ihnen: Neigen Sie – ganz allgemein gesprochen – einer bestimmten Partei zu? Wenn ja, welcher?“. Diese Frage soll langfristige Bindungen an eine Partei messen, nicht aber die Wahlabsicht, was in der Frage durch den Zusatz „obwohl sie auch ab und zu eine andere Partei wählen“ verdeutlicht wird. Ändern sich die Angaben zur Frage der Parteiidentifikation häufig, dann messen wir nicht mehr nur das, was wir messen wollten (langfristige Bindungen), sondern zusätzlich etwas anderes (kurzfristige Präferenzen). Nimmt beispielsweise der Anteil der Befragten mit Parteiidentifikation in Wahljahren deutlich zu, dann ist dies ein Indiz dafür, dass mit dieser Frage auch kurzfristige Präferenzen gemessen werden, da in Wahlkampfzeiten die politische Mobilisierung der Bürger steigt. Zur Überprüfung der Stabilität der Parteiidentifikation sind Informationen über individuelle Veränderungen notwendig, d. h. man benötigt Panel- oder Ereignisdaten (Kapitel 2.4). In Deutschland wird die Parteiidentifikation im Sozio-ökonomischen Panel erhoben. Die Analyse von Schmitt-Beck et al. (2006) deutet auf erhebliche

Fluktuationen in den Parteiidentifikationen der Befragten über einen Zeitraum von 18 Jahren hin. Mit dem SOEP lässt sich leider nicht prüfen, ob die Parteiidentifikation – wie theoretisch erwartet werden kann – stabiler ist als die Wahlabsicht und dieser tatsächlich zeitlich vorausgeht, weil die Wahlabsicht nicht erhoben wird.

An diesem Beispiel lässt sich auch zeigen, **dass die Validität von der Reliabilität abhängt, nicht jedoch umgekehrt die Reliabilität von der Validität**. Unsere Messung kann nämlich sehr zuverlässig sein, wenn etwa kurzfristige Schwankungen in den Parteipräferenzen durch die Frage der Parteiidentifikation exakt registriert und angezeigt würden. Da es sich aber um ein Instrument zur Messung der Parteiidentifikation handelt, glauben wir langfristig stabile Bindungen zu messen. Wir messen also sehr verlässlich etwas, das wir gar nicht messen wollen (kurzfristige Präferenzen). Das heißt: Eine zuverlässige Messung *muss nicht* valide sein.

Ist ein Messinstrument unreliabel, dann kann auch nicht angegeben werden, was gemessen wird. Das heißt: Eine unzuverlässige Messung *kann nicht* valide sein. Reliabilität ist also die notwendige (aber nicht hinreichende) Bedingung für Validität.

Es gibt drei Möglichkeiten, Validität festzustellen:

- Inhaltsvalidität
- Kriteriumsvalidität
- Konstruktvalidität

Das Augenmerk wird bei Prüfung der **Inhaltsvalidität** eines Messinstruments vor allem darauf gerichtet, ob das Messinstrument die zu messende Dimension vollständig erfasst. Bei einem mehrdimensionalen Begriff (Status) muss sichergestellt sein, dass jede Dimension angemessen durch die ausgewählten Items repräsentiert wird. Will man die Bereitschaft zu unkonventioneller politischer Beteiligung messen, dann ist die Frage nach der Beteiligung an genehmigten Demonstrationen sicher kein sehr valides Messinstrument, da diese (in westlichen Demokratien) heute zum „normalen“ Aktionsrepertoire zählen. Will man die Qualität der Lehre messen, dann reicht es nicht aus, lediglich das Verhältnis von Lehrenden zu Studierenden zu erfassen. Ebenso ist die schiere Zahl publizierter Artikel und Bücher alleine kein valider Indikator für die Qualität der Forschung.

Kriteriumsvalidität liegt dann vor, wenn das zu messende Konstrukt anhand eines externen Kriteriums überprüft werden kann. „Extern" bedeutet, dass das Kriterium nicht mit der Messung des Konstruktes im Zusammenhang stehen darf. Beispiel: Wir fragen nach der Wahlabsicht bei der nächsten Bundestagswahl und überprüfen unsere Messung anhand des tatsächlichen Wahlergebnisses. Das Problem besteht darin, dass das Kriterium nur selten so exakt bestimmt werden kann, wie in diesem Beispiel. Wenn es um die Messung solcher Tatbestände wie „Rechtsextremismus" oder „Ausländerfeindlichkeit" geht, ist es schwierig, ein externes Kriterium zu finden, anhand dessen die Messung überprüft werden kann. Außerdem gibt es Probleme, wenn die Feststellung des Kriteriums nur auf einer einzigen Messung beruht. Es ist wenig aufschlussreich, eine in Zweifel stehende Messung anhand eines Kriteriums validieren zu wollen, dessen Validität nicht erwiesen ist. Das obige Beispiel ist deshalb untypisch: Nicht immer hat man die Gelegenheit, eine zeitlich frühere Messung (Wahlabsicht bei der nächsten Bundestagswahl) anhand eines später sehr genau feststellbaren Kriteriums (Wahlergebnis) überprüfen zu können.

Bei der Kriteriumsvalidität werden oftmals *prädiktive* und *gleichzeitige* Validität unterschieden (*predictive* und *concurrent validity*). Beide unterscheiden sich lediglich darin, ob das externe Kriterium später erhoben wird, und sich die Validität dementsprechend in der Vorhersage erweist (*predictive*), oder ob das Kriterium gleichzeitig mit der Messung erhoben wird, und sich die Validität in einer Übereinstimmung zwischen diesen beiden zeigt (*concurrent*). Ein Beispiel für prädiktive Validität haben wir bereits oben angesprochen (Wahlabsicht → Wahlergebnis). Ein Beispiel für gleichzeitige Validität wäre etwa die Messung von „Sympathie für eine Partei" und die Feststellung, ob der Befragte Mitglied dieser Partei ist. Wenn „Parteisympathie" valide gemessen wird, sollte sie bei Mitgliedern höher ausfallen als bei Nichtmitgliedern. Dieses Verfahren wird auch als „Methode der bekannten Gruppen" bezeichnet (vgl. Schnell et al. 2008, 159).

Der Überprüfung der **Konstruktvalidität** liegt der Gedanke zugrunde, dass sich aus den theoretisch begründbaren Beziehungen des zu messenden Konstrukts zu anderen Konstrukten Hypothesen ableiten lassen, die empirisch geprüft werden können. Wenn man gültig misst, müssten sich die durch die Hypothesen behaupteten Zusammenhänge empirisch feststellen lassen. Aus der Rechtsextremismusforschung wissen wir, dass Rechtsextremismus mit Ausländerfeindlichkeit und Nationalismus einhergeht. Ist

die Messung von Rechtsextremismus valide, dann müsste sich ein positiver Zusammenhang zwischen Rechtsextremismus und Ausländerfeindlichkeit und Rechtsextremismus und Nationalismus zeigen. Zeigen sich die vermuteten Zusammenhänge, dann deutet dies darauf hin, dass die Messung valide ist. Zeigen sich die erwarteten Zusammenhänge allerdings nicht, so kann dies mehrere Ursachen haben; unter anderem die, dass eines der Konstrukte nicht valide gemessen wurde oder die Hypothesen falsch sind.

Das „Multitrait-Multimethod"-Verfahren (vgl. Eagly und Chaiken 1993, 69–71) stellt eine besondere Form der Konstruktvalidierung dar, die auf Campbell und Fiske (1959) zurückgeht. Das Verfahren setzt voraus, dass mehrere Konstrukte jeweils durch verschiedene Methoden (z. B. Befragung und Beobachtung) gemessen werden. Die Messungen sollen mit verschiedenen Methoden durchgeführt werden, da hohe Zusammenhänge zwischen verschiedenen Indikatoren eines Konstrukts nicht nur dann auftreten, wenn diese tatsächlich dieselbe Dimension messen, sondern auch durch die Methode bedingt sein können. In einer Umfrage können z. B. ähnliche Stimuli in der Frageformulierung, Zustimmungstendenzen (bei gleich gepolten Items) oder sozial erwünschtes Antwortverhalten für Zusammenhänge zwischen den Items verantwortlich sein (vgl. Kapitel 4.1). Die Verwendung verschiedener Konstrukte dient dem Nachweis, dass die zur Messung der verschiedenen Konstrukte herangezogenen Items tatsächlich Unterschiedliches messen. Unter *Konvergenzvalidität* wird die Stärke des Zusammenhangs zwischen den mit verschiedenen Methoden durchgeführten Messungen eines Konstrukts verstanden; unter *Diskriminanzvalidität* die Stärke des Zusammenhangs zwischen den mit denselben Methoden durchgeführten Messungen verschiedener Konstrukte. Liegt Konstruktvalidität vor, dann sollte die Konvergenzvalidität höher ausfallen als die Diskriminanzvalidität.

In der Praxis wird das „Multitrait-Multimethod"-Verfahren in dieser Form nur selten durchgeführt, weil in der Regel keine auf verschiedenen Methoden beruhenden Messungen vorliegen. Statt dessen kann man das Verfahren so abwandeln, dass zur Messung der Konstrukte verschiedene Indikatoren, die durch eine Methode erhoben wurden, herangezogen werden (vgl. Schnell et al. 2008, 158 ff.).

Aufgaben zu Messen

1. Was bedeutet „Messen"?
2. Nennen Sie die verschiedenen Skalenniveaus und die Eigenschaften, durch die diese charakterisiert werden.
3. Im ALLBUS 1990 wurden unter anderem die unten angegebenen Merkmale (in Klammern: Merkmalsausprägungen) erfasst. Bestimmen Sie bitte das Skalenniveau.
   - Geschlecht (männlich – weiblich)
   - Sind Gewerkschaften für unser Land ...? (hervorragend – sehr gut – gut – nicht besonders gut – überhaupt nicht gut)
   - Alter (in Jahren)
   - Einkommen (in DM)
   - Wahlabsicht (CDU – SPD – FDP – Bündnis 90/Die Grünen – Die Linke – Sonstige)
   - Einkommen (unter 2000 DM – 2001 bis 5000 DM – mehr als 5000 DM)
   - Politisches Interesse (sehr stark – stark – mittel – wenig – überhaupt nicht stark)
   - Religionszugehörigkeit (katholisch – evangelisch – andere)
   - Geburtsjahr
4. Welche der folgenden Antworten ist falsch? Warum?
   a) Intervallskalen haben die Eigenschaften von Nominalskalen.
   b) Ratioskalen haben die Eigenschaften von Ordinalskalen.
   c) Ordinalskalen haben die Eigenschaften von Intervallskalen.
   d) Ordinalskalen haben die Eigenschaften von Nominalskalen.
5. Beschreiben Sie mit eigenen Worten, was Skalierungsverfahren sind. Welche Vorteile haben Skalen im Vergleich zu einem Messinstrument aus einem Indikator?
6. Beschreiben Sie die Konstruktion der behandelten Skalierungsverfahren, und nennen Sie die wesentlichen Vor- und Nachteile.
7. Was bedeuten die Begriffe *Reproduzierbarkeitskoeffizient* und *Item-Analyse*, und in welchem Zusammenhang werden diese gebraucht?

8. Sie haben ein neues Messinstrument zur Messung von Ausländerfeindlichkeit entwickelt. Dieses führt bei wiederholter Anwendung zu stabilen Ergebnissen. Zudem korreliert das neu entwickelte Messinstrument stark mit einer – bereits bewährten – Skala zur Messung von Ausländerfeindlichkeit. Deuten diese Resultate darauf hin, dass Ihr Messinstrument reliabel, valide oder beides ist? Begründen Sie Ihre Antwort.

4 Erhebungsmethoden

In diesem Kapitel geht es darum, auf welche Art und Weise man sich Informationen über einen Ausschnitt der sozialen Realität beschaffen kann. Wollen wir beispielsweise wissen, wie stark fremdenfeindliche Tendenzen in der Polizei vertreten sind, so könnten wir Polizisten befragen (vgl. Mletzko und Weins 1999), wir könnten aber auch das Verhalten von Polizisten gegenüber Ausländern und Deutschen in verschiedenen Situationen (z. B. bei Demonstrationen) beobachten und daraus Rückschlüsse ziehen. Ebenso könnten Strafanzeigen oder Dienstaufsichtsbeschwerden gegen Polizisten auf ihre Ursache (z. B. Diskriminierung von Ausländern) untersuchen. Damit sind die drei zentralen Instrumente der Datengewinnung, die auch als *Erhebungstechniken* oder *Erhebungsmethoden* bezeichnet werden, bereits umrissen:

1. Befragung
2. Beobachtung
3. Inhaltsanalyse

Mit jedem der genannten Erhebungsinstrumente sind bestimmte Vor-, aber auch Nachteile verbunden. Bei einer Befragung von Polizisten zu ihren fremdenfeindlichen Einstellungen kann man nicht sicher sein, ob die Polizisten ihre tatsächlichen Einstellungen berichten oder ob sie diese, z. B. aus Angst vor Sanktionen, lieber verschweigen. Dieses Problem taucht bei einer Beobachtung des *Verhaltens* von Polizisten (z. B. bei Großeinsätzen) nicht auf. Allerdings steht hier der Beobachter vor der schwierigen Aufgabe, alle relevanten Ereignisse gleichzeitig erfassen und einordnen zu müssen.

Insbesondere Politikwissenschaftler und Soziologen stützen sich vorwiegend auf Umfragedaten, weshalb die Befragung im Mittelpunkt dieses Kapitels steht. Die Inhaltsanalyse ist immer noch eine Domäne der Publizistik, da sie sich besonders gut zur Analyse von Medien eignet, während die Beobachtung eher in der Psychologie zu finden ist.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_4,

4.1 Befragung

Weder für den Befragten noch für den Interviewer stellt die Befragung eine natürliche, alltägliche Situation dar. Von einem alltäglichen Gespräch unterscheidet sich ein Interview dadurch, dass eine Person (der Interviewer) nur fragt, während die andere Person – einmal abgesehen von Verständnisfragen – nur antwortet. Das „Gespräch" hat damit einen stark asymmetrischen Charakter. Zudem soll die Befragungsperson einem völlig fremden Menschen zum Teil sehr persönliche Dinge preisgeben. Die Antworten sind also keine an sich schon vorhandenen Informationen, sondern werden durch das Interview erst erzeugt. Aus diesem Grund wird die Befragungssituation auch als *Stimulus-Response-* oder *Reiz-Reaktions-Schema* bezeichnet: Die Frage ist der (künstliche) Reiz, die Antwort die (künstliche) Reaktion. Daher ist die Befragung auch ein reaktives Messverfahren, d. h. die Daten werden durch das Messinstrument – die Befragung – beeinflusst. Mit der Reaktivität von Befragungen sind vor allem zwei methodische Probleme verbunden: die Tendenz von Befragten, sozial erwünscht zu antworten, und die Tendenz, Aussagen unabhängig von ihrem Inhalt zuzustimmen.

Sozial erwünschtes Antwortverhalten („social desirability") liegt vor, wenn die Antwort in Richtung der vermeintlich vom Interviewer bzw. der Gesellschaft als positiv bewerteten Antwort (dem *Ort sozialer Erwünschtheit*) verzerrt ist. Dies wäre z. B. der Fall, wenn ein Befragter vorhandene fremdenfeindliche Einstellungen nicht äußert oder abschwächt, weil er solche Einstellungen als unerwünscht ansieht. Eine Erklärung sieht die Ursache im Streben von Befragten nach sozialer Anerkennung. Bei einer wahren (sozial unerwünschten) Antwort entstehen für den Befragten durch den Verzicht auf soziale Anerkennung Kosten, die er durch sozial als erwünscht angesehene Antworten vermeiden kann (vgl. Reinecke 1991). Antwortverzerrungen durch **Soziale Erwünschtheit** sollten vor allem bei sensiblen Fragen – Fragen, die dem Befragten unangenehm sind – und bei Personen mit einem hohen Bedürfnis nach sozialer Anerkennung stark ausgeprägt sein. Die gesamte Vorurteilsforschung ist mit dem Problem sozialer Erwünschtheit konfrontiert. So wurden in den USA Zweifel an der in Umfragen gemessenen Abnahme von Vorurteilen gegenüber der schwarzen Bevölkerung nach dem Zweiten Weltkrieg laut. Vermutet wurde, dass nicht die Vorurteile abgenommen hatten, sondern lediglich die Äußerung von Vorurteilen zurückgegangen sei, weil die Akzeptanz von Vorurteilen

gegenüber der schwarzen Bevölkerung geringer sei als noch einige Jahrzehnte zuvor (vgl. Dovidio und Gaertner 1986).

Zur Vermeidung oder Abschwächung sozial erwünschten Antwortverhaltens sollten die *Fragen* möglichst *neutral* formuliert sein. Damit versucht man die *Reaktivität des Messinstrumentes* abzuschwächen. In den USA wurden zusätzlich zu den traditionellen Skalen zur Messung von offenen Vorurteilen (blatant prejudice) Skalen entwickelt, die weniger reaktiv sein sollen und auch subtilere Vorurteile erfassen (subtle prejudice). Eine andere Strategie besteht darin, die Anonymität der Befragungssituation zu erhöhen. Im ALLBUS 2006 wurde die Anonymität der Befragung bei einigen Items zur Einstellung gegenüber Ausländern (vgl. Abbildung 3.2, S. 49) für einen zufällig ausgewählten Teil der Befragten erhöht. Die Antworten wurden bei diesen Befragten nicht vom Interviewer erhoben. Vielmehr haben die Befragten ihre Antworten selbst (und für den Interviewer nicht kontrollierbar) im Computer eingegeben (CASI-Split, vgl. S. 25). Man sollte erwarten, dass soziale Erwünschtheit bei eigener Eingabe der Antworten eine geringere Rolle spielt als bei Abfrage durch den Interviewer. Außerdem kann man soziale Erwünschtheit direkt über Skalen erfassen (vgl. Reinecke 1991). Die Skalen beruhen auf der Annahme, dass sozial erwünschtes Antwortverhalten aus dem Bedürfnis der Befragten nach sozialer Anerkennung resultiert. Im ALLBUS 2006 wurde eine Kurzform der deutschen Fassung der Marlowe-Crowne-Skala mit zehn Items verwendet. Der Einfluss des Bedürfnisses nach sozialer Anerkennung auf das Antwortverhalten kann mit diesen Skalen direkt geprüft werden.

Mit **Zustimmungstendenz** (Akquieszenz) oder „Ja-Sage-Tendenz“ wird ein Verhalten von Befragten beschrieben, einer Frage unabhängig von ihrem Inhalt zuzustimmen. Zustimmungstendenzen lassen sich prüfen, wenn mehrere, negativ und positiv formulierte Items zur Messung einer Dimension verwandt werden. Befragte, die negativ formulierten Items zustimmen, sollten positiv formulierte Items ablehnen und umgekehrt. Befragte, die sich unabhängig von der Polung der Frage immer zustimmend äußern, antworten unplausibel und können gegebenenfalls aus der Analyse ausgeschlossen werden. Die Zustimmungstendenz ist ein Spezialfall eines *response rets*. *Reponse sets* bezeichnen allgemein die Neigung von Befragten, Items unabhängig von der zu messenden Dimension in einer bestimmten Art und Weise zu beantworten. Dazu zählen die Bevorzugung oder Vermeidung der Mittelkategorie ebenso wie eine Präferenz für die Extremka-

tegorien. Zur Kontrolle von Response-Sets können in einem ersten Schritt die Antwortmuster der Befragten ausgewertet werden.

Selbst wenn der Befragte gewillt ist, korrekt Auskunft zu geben, lassen sich Messfehler nicht gänzlich vermeiden, z. B. aufgrund von Erinnerungsfehlern. Allgemein kann man davon ausgehen, dass Fragen zu demographischen und biographischen Merkmalen (Geschlecht, Ausbildung, Eintritt in eine Partei, Heirat usw.) korrekter beantwortet werden als Fragen zum Verhalten (z. B. Teilnahme an Demonstrationen) und diese wiederum korrekter als Fragen zu Einstellungen (Meinung zum Schwangerschaftsabbruch, Parteiidentifikation usw.), weil demographische und biographische Fakten dem Befragten selbst eher bewusst und weniger flüchtig sind als Verhalten und vor allem Einstellungen. Eine retrospektive Frage nach dem Jahr der Eheschließung dürften die meisten Verheirateten einigermaßen korrekt beantworten. Es ist daher nicht sinnvoll, weit zurückliegende Einstellungen mit retrospektiven Fragen zu erfassen.

4.1.1 Formen der Befragung

Wenn wir von Befragung sprechen, meinen wir in der Regel die *standardisierte* bzw. *quantitative Befragung*. In ihr ist der Verlauf des Interviews durch die exakte Formulierung und genaue Abfolge der Fragen festgelegt. Abweichungen davon sind nicht zulässig. Sind die Fragen und/oder der Ablauf der Befragung nicht fixiert, dann spricht man von einer *nicht-standardisierten* bzw. *qualitativen Befragung* (vgl. Bortz und Döring 2006, 308-321). Die Grenzen zwischen beiden Formen sind fließend. Es kann sich bei einer nicht-standardisierten Befragung z. B. um ein *Leitfadengespräch* handeln, bei dem der Interviewer nur eine Liste von Themen hat, die er in beliebiger Reihenfolge abarbeiten kann. Bei *narrativen Interviews* wird den Befragten lediglich eine Themenstellung vorgegeben. Im folgenden beschäftigen wir uns mit der standardisierten Befragung.

Eine Befragung kann persönlich, schriftlich, telefonisch und internetgestützt erfolgen. Bei der *persönlichen Befragung* besucht ein Interviewer die Befragungsperson und führt mit dieser das Interview durch. Der Interviewer liest der Befragungsperson die Fragen aus dem Fragebogen vor und notiert die Antworten. Mit Ausnahme von Kärtchen und anderen visuellen Hilfen, die der Veranschaulichung dienen, bekommt die Befragungsperson nichts vorgelegt. Eine Vielzahl von Beispielen für visuelle Hilfen findet man bei Noelle-Neumann und Petersen (1996). Bei der *schriftlichen Befragung*

wird der Befragungsperson ein Fragebogen zum Selbstausfüllen überreicht oder zugestellt. Üblicherweise soll die Befragungsperson den Fragebogen zurücksenden, in Ausnahmefällen wird der Bogen auch abgeholt. Befragt man ganze Schulklassen in Klassenräumen, dann können die Bögen direkt überreicht und auch unmittelbar nach der Beantwortung der Fragen wieder eingesammelt werden. Die *telefonische Befragung* läuft ähnlich wie eine persönliche Befragung ab. Allerdings können dem Befragten keine visuellen Hilfen gegeben werden. Bei internetgestützten Befragungen (vgl. Janetzko 1999; Batinic et al. 1999) füllt der Befragte entweder direkt im Internet auf einem Web-Server einen Fragebogen aus (*Web-Survey*) oder er bekommt diesen per E-Mail zugesandt.

Persönliche Befragungen haben ihren dominanten Stellenwert in der Markt- und Meinungsforschung in den vergangenen zehn Jahren eingebüßt. Auch schriftliche Befragungen sind hier seit 1990 deutlich zurückgegangen (Tabelle 4.1).[1] Telefonumfragen (vgl. Frey et al. 1990; Fuchs 1994) sind inzwischen am weitesten verbreitet. Besonders dynamisch entwickeln sich Online-Interviews, die vor zehn Jahren noch keine Bedeutung hatten, inzwischen aber fast ein Drittel aller Interviews der Mitgliederinstitute des ADM ausmachen. Berücksichtigt muss bei diesen Angaben allerdings, dass Online-Surveys vor allem in der Marktforschung eingesetzt werden.

Tabelle 4.1: Formen der Befragung in der Markt- und Meinungsforschung

	1990	1995	2000	2005	2007
Persönliche Interviews	65 %	60 %	34 %	24 %	26 %
Telefoninterviews	22 %	30 %	41 %	45 %	41 %
Schriftliche Interviews	13 %	10 %	22 %	9 %	6 %
Online-Interviews	-	-	3 %	22 %	27 %

Quelle: Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (2006, 16) und `www.adm-ev.de`

Ein Grund für diese Entwicklung ist in der Kostenstruktur zu sehen. Persönliche Interviews sind deutlich teurer als telefonische Befragungen, weil die Interviewer vor Ort die Interviews durchführen, wodurch Reisekosten usw. anfallen. Telefoninterviews können dagegen zentral in einem Telefonlabor durchgeführt werden. Durch die 1998 erfolgte Liberalisierung

1 Aufgrund von Veränderungen bei den Mitgliedsinstituten des ADM sind die Zahlen über die Jahre nicht unmittelbar vergleichbar.

des Telekommunikationsmarktes sind die Kosten für Telefonate im Festnetz und mit dem Handy zudem dramatisch gesunken. Die dynamische Entwicklung bei Online-Interviews wurde erst durch die Verbreitung von Computern und Internetanschlüssen in privaten Haushalten möglich. Die abnehmende Bedeutung schriftlicher Befragungen wird auf eine Ersetzung durch Online-Erhebungen zurückgeführt (vgl. Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. 2006).

Der Reiz internetbasierter Umfragen liegt vor allem in den geringen Kosten und der schnellen Verfügbarkeit der Daten. Im Gegensatz zu telefonischen Interviews können visuelle Hilfen jeder Art (auch Filme) verwendet werden. Internetbasierte Umfragen sind jedoch mit besonderen Problemen behaftet (vgl. für eine kritische Einschätzung Schnell et al. 2008, 379 f.). Dazu zählt vor allem die Schwierigkeit der Realisierung einer Zufallsstichprobe (vgl. Kapitel 9), die für die Verallgemeinerung von Stichprobenergebnissen notwendig ist. Bevölkerungsweite Umfragen sind derzeit mit internetgestützten Befragungen nicht möglich, weil lediglich die Hälfte der bundesdeutschen Bevölkerung das Internet privat nutzt (Angaben nach ALLBUS 2006) und sich die Internetnutzer von den Nicht-Nutzern in wesentlichen Merkmalen unterscheiden - sie sind beispielsweise jünger. Für bestimmte Fragestellungen und Zielpopulationen können Internet-Befragungen jedoch sinnvoll sein.

Die technologische Revolution durch Mikrocomputer hat auch telefonische und mündliche Befragungen erfasst. In computergestützten Telefonbefragungen (CATI – *Computer Assisted Telephone Interview*) wird der Fragebogen programmiert. Der Interviewer liest die Fragen während des Interviews vom Bildschirm ab und gibt die Antworten der Befragten direkt in den Computer ein. Automatisch wird dann zur nächsten Frage gesprungen, wobei das Programm automatisch für die richtige Filterführung sorgt. Auch mündliche Befragungen können so durchgeführt werden (CAPI – *Computer Assisted Personal Interview*). Die Interviewer müssen dazu mit einem tragbaren Computer ausgestattet werden. Die Verwendung von Computern bei der Datenerhebung bietet unbestreitbare Vorteile. Es können beispielsweise komplexe Filterführungen (siehe unten) eingesetzt werden, weil der Computer automatisch zur nächsten Frage springt. Eine vom Interview getrennte Erfassung der Daten entfällt, weil die Dateneingabe während des Interviews statt findet. CATI und CAPI sind mehr als programmierbare Fragebögen. CATI-Systeme verwalten beispielsweise Telefonnummern und führen die Anrufe (zur gewünschten Zeit) aus. Te-

lefonische Erhebungen werden inzwischen weitgehend computergestützt durchgeführt. Der Anteil der computergestützten persönlichen Befragungen (CAPI) hat sich zwischen 2000 und 2006 bei den Mitgliedsinstituten des ADM von ca. einem Viertel auf knapp die Hälfte verdoppelt (vgl. Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. 2006, 15; Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. 2000, 10).

Unabhängig von der Form der Befragung müssen bestimmte Grundprinzipien beachtet werden. Im schlimmsten Fall vergisst man eine Frage zu stellen, die für den Untersuchungszweck relevant ist. Nachträglich kann ein solcher Fehler nicht mehr behoben werden. Man sollte sich deshalb vor der Entwicklung des Fragebogens nicht nur darüber im Klaren sein, was man erklären will (z. B. fremdenfeindliche Einstellungen von Polizisten), sondern sich auch genau überlegen, welche Faktoren die abhängige Variable beeinflussen könnten (z. B. dienstliche Belastungen, das Einsatzgebiet, die politische Einstellung, das Alter oder das Geschlecht der Polizeibeamten). Weiß man, welche Aspekte erfasst werden sollen, so kann man sich an die Formulierung der Fragen wagen und schließlich den Fragebogen zusammenstellen. Bei der Formulierung der Fragen sollte man die im Folgenden genannten Aspekte berücksichtigen (vgl. Converse und Presser 1986; Sudman 1982).

4.1.2 Die Fragen

Die *Frageformulierungen* und die *Antwortmöglichkeiten* prägen das Antwortverhalten der Befragten entscheidend (vgl. Schuman und Presser 1996). Aus diesem Grund sollte man sich bei der Interpretation von Umfrageergebnissen nicht nur die Verteilung der Antworten, sondern auch die gestellten Fragen genau anschauen. So stimmten in einer im Herbst 1995 durchgeführten Umfrage unter den Mitgliedern der rheinland-pfälzischen SPD erstaunliche 74 % der Befragten einem Lauschangriff zu (Kategorien „stimme voll und ganz zu" und „stimme eher zu"), wie man einer Ergebnisdokumentation entnehmen kann. Dieses Ergebnis ist weit weniger erstaunlich, wenn man die – in der erwähnten Ergebnisdokumentation nicht berichtete – Frageformulierung berücksichtigt. Sie lautete: „Bei der Verfolgung besonders schwerer Straftaten soll das rechtsstaatlich geregelte Abhören des gesprochenen Wortes zu Beweiszwecken verwendet werden dürfen (sog. Lauschangriff)." Die Einschränkung auf „besonders schwere

Straftaten“ und der Verweis auf die Rechtsstaatlichkeit des Verfahrens trug sicher zu dieser hohen Zustimmung bei.

Bei den meisten Fragen kann man die Antworten in eine wie auch immer gewünschte Richtung beeinflussen. In manchen Fällen scheinen die „Frager“ weniger an den tatsächlichen Meinungen, Einstellungen oder Verhaltensweisen der befragten Personen interessiert zu sein als an einer Bestätigung bestimmter Positionen. Die Ergebnisse einer Umfrage können dann z. B. in der politischen Auseinandersetzung zur Stützung des eigenen Standpunktes herangezogen werden (vgl. Diekmann 2008, 458 f.).

Regeln zur Frageformulierung

In wissenschaftlichen Umfragen spielen bewusste Manipulationen (in der Regel) keine Rolle. Bei der Publikation wissenschaftlicher Ergebnisse auf Basis von Umfragedaten werden die Formulierungen der Fragen häufig mit veröffentlicht. Durch offensichtlich einseitig gestellte Fragen würde der Wissenschaftler die Glaubwürdigkeit seiner Ergebnisse von Vornherein in Frage stellen. Unbeabsichtigt schleichen sich dagegen in jeder Umfrage Formulierungen ein, die sich auf das Antwortverhalten auswirken. Hält man sich an einige Grundregeln, kann man jedoch schwerwiegende Fehler vermeiden.

Generell sollten die Fragen so formuliert sein, dass sie den Befragten nicht überfordern. Das heißt:

- kurze Fragen stellen,
- einfache und allgemeinverständliche Begriffe und Formulierungen verwenden,
- konkrete Dinge ansprechen,
- eindeutige Begriffe benutzen,
- (doppelte) Verneinungen (Negationen) vermeiden,
- keine Suggestivfragen stellen, und
- keine mehrdimensionalen Fragen verwenden.

In der Frage sollten – wenn möglich – bereits alle Antwortalternativen „ausformuliert“ sein, damit keine Antwortalternative durch die Nennung in der Frage bevorzugt wird. Solche Fragen werden auch als „balancierte“

Fragen bezeichnet. Es lässt sich nämlich zeigen, dass die Antwortalternative, die in der Frage enthalten ist, deutlich mehr Zustimmung erhält als eine ungenannte Antwortalternative (vgl. die Beispiele bei Noelle-Neumann und Petersen 1996, 131 f. und 195 ff.). Balancierte Fragen sind natürlich nur bei einer geringen Zahl von Antwortalternativen möglich. Will man die Zustimmung oder Ablehnung zu *einer* Position messen, wie es z. B. bei Likert-Skalen der Fall ist, dann können keine „balancierten" Formulierungen verwendet werden.

Man kann vermuten, dass die oben genannten Antworten der rheinland-pfälzischen SPD-Mitglieder zum „großen Lauschangriff" anders ausgefallen wären, wenn eine balancierte Frage formuliert worden wäre, etwa in der Art: „Sind Sie der Meinung, dass bei der Verfolgung besonders schwerer Straftaten das Abhören des gesprochenen Wortes zu Beweiszwecken verwendet werden darf oder sind Sie der Meinung, dass bei der Verfolgung besonders schwerer Straftaten das Abhören des gesprochenen Wortes *nicht* verwendet werden darf". Als Antwortkategorien könnte man vorgeben „sollte nicht verwendet werden" und „sollte verwendet werden". Eine Skala von „stimme stark zu" bis „lehne stark ab" ist hier nicht mehr möglich, da die Frage ja mehrere Positionen enthält. Darauf, dass die Schärfe der Formulierung („besonders schwere Straftaten") das Antwortverhalten beeinflusst, wurde bereits hingewiesen.

Generell gilt, dass die Fragen so *kurz* wie möglich sein sollten, damit der Befragte am Ende des Satzes auch noch weiß, worum es geht. Dies ist bei mündlichen und telefonisch durchgeführten Interviews wichtiger als bei schriftlichen Befragungen, in denen der Befragte die Möglichkeit hat, sich die Frage mehrmals durchzulesen.

Wichtig ist bei bevölkerungsweiten Umfragen auch, dass man einfache und *allgemein verständliche Begriffe* verwendet und Fremdwörter vermeidet, da man nicht bei allen Befragten von einem gleich großen Wortschatz ausgehen kann: Statt „partiell" schreibt man also besser „teilweise", und verständlicher als „Applikation" ist sicher „Anwendung". Versteht ein Befragter die Frage nicht, so besteht die Gefahr, dass er einfach irgendetwas ankreuzt oder die Antwort verweigert. Vor dieser Unannehmlichkeit kann man die Befragten in der Regel ohne größere Probleme bewahren. Die Sprache sollte der Zielgruppe der Untersuchung angemessen sein.

Schon schwieriger ist die Forderung, nur nach *konkreten Dingen* zu fragen, d. h. abstrakte Begriffe zu vermeiden. In Umfragen findet sich z. B.

häufig die Frage „Wie zufrieden sind Sie mit der Demokratie?“. Hier kann man den Antworten nicht viel abgewinnen, da man nicht so genau weiß, was der Befragte unter Demokratie versteht (z. B. das politische System der Bundesrepublik?). Ein weiteres Beispiel ist die Frage „Wie stark ist ihr politisches Interesse?“. Diese Frage ist nicht *eindeutig*. Hier muß der Befragte entscheiden, woran er sein politisches Interesse festmacht; an der Häufigkeit, mit der er den politischen Teil der Tageszeitung liest? Ob er sich am politischen Leben, z. B. in einer Bürgerinitiative oder einer politischen Partei beteiligt? Es wäre sinnvoller, mehrere Fragen zu stellen, die genau diese Aspekte als Indikatoren für politisches Interesse erfassen. Damit wäre der Befragte nicht auf seine eigene Interpretation angewiesen, die der Forscher aus der gegebenen Antwort nicht mehr erkennen kann. Auch *Verneinungen* in der Frage tragen nicht zu einer besseren Verständlichkeit bei, weil eine ablehnende Antwort zu einer doppelten Verneinung führt: „Es ist nicht die Aufgabe der Opposition, die Regierung zu kritisieren, sondern sie in ihrer Arbeit zu unterstützen“.

Suggestivfragen, d. h. Fragen, durch deren Formulierung den Befragten eine bestimmte Antwort nahegelegt wird, gehören in den Bereich der Manipulation und sind schlicht unzulässig. Eine Suggestivfrage wäre etwa „Sind Sie auch der Meinung, dass die Ausgaben für BAföG viel zu hoch sind?“. Die Frage „Sind Sie für die Verringerung von Treibhausgasen und eine längere Laufzeit von Atomkraftwerken?“ ist aus einem anderen Grund falsch. Hier handelt es sich um eine *mehrdimensionale Frage*, d. h. eine Frage, die verschiedene Aspekte beinhaltet. Diese verschiedenen Aspekte können nicht getrennt beantwortet werden, weil sie in einer Frage zusammengefasst sind. Im Beispiel zwingen wir den Befragten, entweder die Reduktion von Treibhausgasen *und* längere Laufzeiten zu befürworten oder beides abzulehnen. Dem Befragten wird damit die Möglichkeit genommen, sich für die Klimaziele und gegen längere Laufzeiten zu äußern und umgekehrt. Zwei Dimensionen in einer Frage sind deshalb unbedingt zu vermeiden.

Antwortformate

Entscheidend für die Ergebnisse einer Umfrage ist auch, ob *Antwortalternativen* vorgegeben werden (*geschlossene Fragen*) oder die Befragten in einem dafür vorgesehenen freien Feld eine Antwort niederschreiben können (*offene Fragen*).

Üblicherweise wird man geschlossene Fragen dort einsetzen, wo die Antwortmöglichkeiten *bekannt* und *begrenzt* sind. Um geschlossene Kategorien vorzugeben, muss man also schon wissen, welche Antworten gegeben werden können. Das einfachste Beispiel ist hier das Merkmal Geschlecht mit den Ausprägungen männlich und weiblich. Offene Fragen gibt man den Vorzug, wenn man noch keine Vorstellung davon hat, was die Befragten antworten werden; die Kenntnis über den Untersuchungsgegenstand also noch ziemlich gering ist. Offene Fragen sind auch dann geeigneter als geschlossene, wenn die Antwortmöglichkeiten unbegrenzt oder zumindest sehr vielfältig sind. Um beispielsweise das Einkommen oder das Alter hinreichend genau zu erfassen, müssten sehr viele Kategorien verwendet werden, während bei einer offenen Frage nur ein Feld notwendig ist.

Auch bei anderen Fragen, wie etwa nach der Zahl der am Vortag gerauchten Zigaretten, sollte man sich überlegen, ob man nicht die genaue Zahl offen erfasst. Die Kategorisierung selbst beeinflusst nämlich das Antwortverhalten. Bei niedrigen Antwortvorgaben `[0 | 1-4 | 5-9 | 10-19 | 20 und mehr]` wird man wahrscheinlich einen niedrigeren Zigarettenkonsum ermitteln als bei hohen Antwortvorgaben `[0 | 1-19 | 20-29 | 30-39 | 40 und mehr]`, weil die Kategorisierung Anhaltspunkte für die Einschätzung des eigenen Verhaltens liefert (vgl. Schwarz et al. 1985).

Auch die Reihenfolge der Antwortvorgaben und deren Visualisierung kann das Antwortverhalten beeinflussen (Beispiele finden sich bei Noelle-Neumann und Petersen 1996, 191–207). Reihenfolgeeffekte lassen sich kontrollieren, in dem die Abfolge der Antwortkategorien zufällig variiert wird. Dies ist natürlich nur bei *ungeordneten* Antwortkategorien (nominale Merkmale) sinnvoll.

Einen Kompromiss zwischen geschlossener und offener Frage stellt die *Hybridfrage* dar. Hier hat der Befragte die Möglichkeit, eine vorgegebene Kategorie anzukreuzen. Trifft keine dieser Kategorien zu, so kann der Befragte in einem Feld „Sonstiges“ offen antworten. Die Frage nach der Wahlabsicht, aufgrund der üblicherweise verwendeten Formulierung auch „Sonntagsfrage“ genannt, ist ein Beispiel für eine Hybridfrage (vgl. Abbildung 4.1). Die Antwortmöglichkeiten bei der Wahlabsicht sind begrenzt und bekannt – nämlich die kandidierenden Parteien. Da bei einer Bundestagswahl aber schon einmal 20 oder mehr Parteien antreten, wäre es übertrieben, alle aufzulisten, zumal manche Parteien selten oder nie ge-

nannt würden. Man gibt also nur die Parteien vor, die meistens genannt werden und lässt die Möglichkeit zu, neben dem Feld „andere Partei“ eine nicht aufgeführte Partei zu nennen.

Abbildung 4.1: Sonntagsfrage im ALLBUS 1990

S 70A	*Liste S70A vorlegen!*		26/27	
	Wenn am nächsten Sonntag Bundestagswahl wäre, welche Partei würden Sie dann mit Ihrer **Zweitstimme** wählen?	CDU bzw. CSU	01	
		SPD	02	
		F.D.P.	03	
		Die Grünen	06	
	Nur eine Nennung möglich!	NPD	04	
		DKP	05	S 71
		Die Republikaner	07	
		Andere Partei *(bitte notieren)*: ________	08	
		Würde nicht wählen	10	
		Angabe verweigert	97	
		Weiß nicht	98	
			99	
S 70B	*Liste S70B vorlegen!*		28/29	
	Wenn am nächsten Sonntag Wahl zum Berliner Abgeordnetenhaus wäre, welche Partei würden Sie dann mit Ihrer **Zweitstimme** wählen?	CDU bzw. CSU	01	
		SPD	02	
		F.D.P.	03	
		Alternative Liste	06	
	Nur eine Nennung möglich!	SEW	05	
		Die Republikaner	07	
		Andere Partei *(bitte notieren)*: ________	08	
		Würde nicht wählen	10	
		Angabe verweigert	97	
		Weiß nicht	98	
			99	

Geschlossene Fragen haben ganz allgemein den Vorteil, dass durch die Standardisierung der Antworten die Auswertung der Fragen erleichtert wird. Die offenen Antworten müssen zunächst alle erhoben werden, bevor ähnliche Antworten zu Gruppen zusammen gefasst werden können. Erst nach diesem Prozess der *Kategorisierung* kann die Auswertung beginnen. Dies ist ein zeit- und kostenaufwändiges Verfahren, vor allem bei großen Umfragen. Zudem ist die Vergleichbarkeit der Antworten bei geschlosse-

nen Fragen höher als bei offenen. Mit dem Verzicht auf offene Fragen ist allerdings immer ein Informationsverlust verbunden.

Entscheidet man sich für geschlossene Fragen, so muss man zwei Dinge beachten: Die Antwortkategorien müssen die *Bandbreite möglicher Antworten erschöpfend abdecken*. Zudem müssen sich die einzelnen Antwortmöglichkeiten *gegenseitig ausschließen*. Die erste Forderung nach Vollständigkeit kann man erfüllen, indem man im Zweifelsfall eine Restkategorie „Sonstiges“, „Andere“ etc. vorgibt. Ebenso ist die Forderung nach Ausschließlichkeit selbstverständlich: Ein Befragter darf sich nur in einer der vorgegebenen Kategorien wiederfinden, d. h. die Antwortmöglichkeiten dürfen sich nicht überlappen.

Bei *Mehrfachantworten* wird den Befragten die Möglichkeit eingeräumt, mehrere Kategorien anzukreuzen. Dies steht nur scheinbar im Widerspruch zum Prinzip, dass sich die Antwortalternativen gegenseitig ausschließen müssen, da hier mehrere Fragen in einer Frage (meist aus Platzgründen) zusammengefasst werden. Mit der in Abbildung 4.2 dargestellten Frage werden z. B. Vereinsmitgliedschaften ermittelt. Hier sind mehrere Nennungen möglich. Kreuzt ein Befragter ein Kästchen an, dann bedeutet dies, dass er „Mitglied“ in einem bestimmtem Vereinstyp ist; wird nichts angekreuzt, heißt dies, der Befragte ist „kein Mitglied“. Diese beiden Antwortmöglichkeiten für jeden genannten Punkt (hier: Vereinstyp) schließen sich also gegenseitig aus. Die einzelnen Antworten müssen jeweils als eigene Variable kodiert werden. Als Ergebnis dieser Frage erhält man fünf Variablen mit jeweils zwei Antwortmöglichkeiten. Schnell et al. (2008, 333 f.) schlagen zur Fehlervermeidung vor, jeweils beide Antwortmöglichkeiten vorzugeben. In Abbildung 4.2 müssten pro Verein also jeweils zwei Kästchen vorgegeben werden, die mit „Mitglied“ bzw. „kein Mitglied“ beschriftet werden müssten.

Eine besondere Form geschlossener Antwortkategorien sind Rating-Formate, wie sie auch für die bereits mehrfach verwendeten Aussagen zur Messung ausländerfeindlicher Einstellungen im ALLBUS 2006 verwendet wurden (Abbildung 4.3). Bei Rating-Formaten können Befragte ihre Beurteilung in abgestufter Form vornehmen (*geordnete* Antwortalternativen). Bei den ALLBUS-Items wurde eine siebenstufige, *bipolare* Skala von *stimme überhaupt nicht zu* – 1 – bis *stimme voll und ganz zu* – 7 – verwendet. Mit der Zuordnung der Zahlen wird bezweckt, dass die Abstände zwischen den Skalenpunkten von den Befragten als gleich

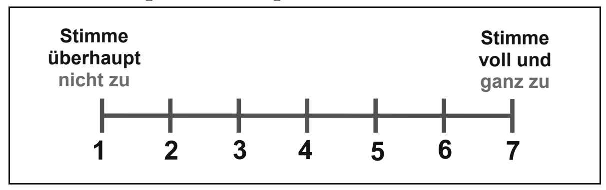
Abbildung 4.2: Frage mit Mehrfachantworten

Sind Sie Mitglied eines Vereins? (Mehrfachnennungen möglich)

Gesangverein	☐
Sportverein	☐
Heimatverein	☐
Caritativer Verein	☐
Anderer Verein	☐

interpretiert werden und die Items damit auf Intervallskalenniveau messen. Diese Annahme ist prüfbar (vgl. Rost 2004). Häufig werden auch fünfstufige Antwortalternativen verwendet. Die Zahl der Antwortstufen hängt davon ab, in welcher Differenziertheit man von den Befragten eine Beurteilung erwarten kann. Die vorhandene Präferenz für Rating-Formate in Umfragen hängt auch damit zusammen, dass mit der Unterstellung gleicher Abstände zwischen den Skalenpunkten die Auswertung der Daten vereinfacht wird.

Abbildung 4.3: Rating-Format mit sieben Stufen



Diskutiert wird, ob man eine gerade oder - wie im Beispiel – ungerade Zahl von Antwortkategorien vorgibt. Bei gerader Zahl der Kategorien existiert keine *mittlere Position*, wodurch die Befragten zu einer Positionierung in Richtung einer der beiden Enden der Skala gezwungen werden. Für die Vorgabe einer mittleren Kategorie spricht, dass man kognitiv sehr wohl eine neutrale Position einnehmen kann. Nachteilig wirkt es sich allerdings aus, wenn Befragte die mittlere Kategorie wählen, um auszudrücken, dass sie keine Position zu diesem Item einnehmen. In diesem Fall misst man Pseudo-Meinungen (*pseudo-opinions*). Durch die explizite Vorgabe einer „Weiß-nicht“-Kategorie (vgl. das Beispiel in Abbildung 4.1) für Mei-

nungslosigkeit (*non-attitude*), kann man dies verhindern. Die „Weiß-nicht“-Kategorie sollte bei mehreren geordneten Antwortalternativen auf keinen Fall als Mittelkategorie verwendet werden. Ohne getrennte Erfassung der „Weiß-nicht“-Antworten kann man bei der Auswertung der Daten auch nicht mehr feststellen, ob eine fehlende inhaltliche Angabe auf Meinungslosigkeit (*non-attitude*) oder Antwortverweigerung (*item-non-response*) beruht. Es ist natürlich auch möglich, Befragte ohne Meinung zu filtern. Zunächst wird dann gefragt, ob eine Meinung zu einem konkreten Thema vorhanden ist. Lediglich den Befragten mit einer Meinung werden die Antworten vorgelegt. Eine Filterführung für Meinungslosigkeit zeigt den Befragten die Legitimität einer solchen Beantwortung noch deutlicher als eine Antwortkategorie „weiß nicht“ (vgl. Schnell et al. 2008, 337). Andererseits muss befürchtet werden, dass Meinungslosigkeit durch einen Filter überschätzt wird. Der Filter selbst mag als Hinweis auf eine schwierige folgende Frage interpretiert werden und auch Befragte, die eine Meinung haben, zu einer „Weiß-nicht“-Antwort bewegen.

Die Wahl der Antwortkategorien bestimmt das Messniveau und damit die zulässigen Auswertungsverfahren. Gibt man auf die Frage „Sollte man Ihrer Meinung nach die doppelte Staatsbürgerschaft erlauben oder sollte man die doppelte Staatsbürgerschaft nicht erlauben?“ die Antwortalternativen „sollte man erlauben“, „sollte man nicht erlauben“ und „weiß nicht“ vor, dann ist die Variable nominal skaliert. Man hätte aber auch eine fünfstufige Antwortskala von „stimme voll und ganz zu“ bis „lehne voll und ganz ab“ als Antwortalternative wählen können, wobei die Frage dann nur noch eine der beiden Positionen enthalten kann, etwa: „Die doppelte Staatsbürgerschaft sollte erlaubt werden“. Dieses Merkmal wäre ordinal skaliert.

4.1.3 Der Fragebogen

Nach der Formulierung der einzelnen Fragen muss über deren Anordnung im Fragebogen nachgedacht werden. Nicht selten werden Interviews von Befragten abgebrochen, weil sie sich scheinbar endlos hinziehen oder zuviel Konzentration erfordern.

Die Befragung wird meist durch so genannte *Aufwärmfragen* eingeleitet, mit denen man die Neugier und das Interesse des Befragten für das Thema der Befragung wecken will. Aus diesem Grunde werden Fragen zur Demographie (Alter, Geschlecht usw.) in der Regel *nicht* zu Beginn der

Befragung gestellt. Stellt man demographische Angaben an den Anfang, dann riskiert man bei Befragten, die an der Anonymität ihrer Angaben zweifeln, eine Verweigerung des Interviews. Durch die Fragen zu ihrer Person können diese in der Befürchtung bestärkt werden, dass sie identifiziert werden könnten. Es kann jedoch auch sinnvoll sein, die Fragen zum eigentlichen Untersuchungsgegenstand nicht zu Beginn eines Interviews zu stellen, wenn diese unangenehme oder schwierige Sachverhalte betreffen. In diesem Fall kann man mit einer anderen, einfachen inhaltlichen Frage beginnen. Zu Anfang der Befragung gilt es, den Befragten für das Interview zu gewinnen, wobei die erste Frage von ausschlaggebender Bedeutung sein kann.

Eine Grundregel für die Fragenanordnung bei Mehrthemenumfragen besteht darin, Fragen zu einem *Themenkomplex* zusammenhängend zu stellen. Auch die Fragen zur Demographie können in einem Block – meist am Ende des Interviews – gestellt werden. Gelegentlich ist die Trennung von Fragen zum selben Thema sinnvoll, um *Halo-Effekte* zu vermeiden. Darunter wird die unerwünschte Ausstrahlung einer Frage auf die nachfolgende Frage verstanden. So könnte z. B. die Frage, ob man für die Todesstrafe sei, *nach* mehreren Fragen zur Kriminalität in der Gesellschaft höhere Zustimmungswerte liefern, als wenn die Frage *davor* gestellt wird. Im ALLBUS 1990 wurde die Reihenfolge der Themenblöcke so festgelegt: Politik/Gesellschaft, AIDS, Soziale Normen, Deutsche Einheit, Demographie, Gesellschaft und zuletzt noch ein paar statistische Angaben zum Interview, unter die die Sonntagsfrage gemischt wurde. Die Wahlabsichtsfrage wurde aus dem Themenblock Gesellschaft/Politik herausgenommen, um eine Beeinflussung durch zuvor gestellte Fragen möglichst gering zu halten.

Ebenso sollten schwierig zu beantwortende Passagen mit einfachen Passagen abwechseln, um die Konzentrationsfähigkeit der Befragten nicht zu sehr zu strapazieren. Aus diesem Grund sollte das Interview auch nicht zu lang sein. Allerdings können keine einheitlichen Angaben zur vertretbaren *Länge eines Interviews* gemacht werden. Die vertretbare Dauer des Interviews hängt von der Relevanz der abgefragten Themen für die Befragten und der Zielgruppe der Befragung ab. Nach Angaben von Hanefeld (1987, 235–238) wurden bei der ersten Welle des SOEP in der Stichprobe A „Deutsche“ zur Beantwortung des Haushaltsfragebogens durchschnittlich knapp 20 Minuten und zur Beantwortung des Personenfragebogens durchschnittlich rund 35 Minuten benötigt. In Tabelle 4.2 ist die Inter-

viewdauer für den ALLBUS 1994 und den ALLBUS 1998 wiedergegeben. Mehr als 35 % der Interviews dauerten jeweils länger als eine Stunde.

Tabelle 4.2: Interviewdauer bei ALLBUS-Umfragen

Minuten	1994	1998
20 bis 39	17 %	9 %
40 bis 59	44 %	54 %
60 bis 74	22 %	25 %
75 bis 99	12 %	9 %
mehr als 100	2 %	1 %
keine Angabe	4 %	2 %
	100 %	100 %
(Befragte)	(3450)	(3234)

Filterfragen dienen dazu, das Interview abzukürzen, wenn auf den Befragten bestimmte Fragen nicht zutreffen. Fragen zum Ehepartner können beispielsweise übersprungen werden, wenn der Befragte nicht verheiratet ist. Ein Beispiel für eine Filterfrage ist die Parteiidentifikation (vgl. Abbildung 4.4), mit der eine langfristig stabile psychologische Bindung an eine Partei gemessen werden soll. Der Befragte soll zunächst angeben, ob er überhaupt einer Partei zuneigt (Frage 30 in Abbildung 4.4). Wie man anhand der Angaben für den Interviewer am rechten Rand des Fragebogens sehen kann, geht der Befragte bei einer „Ja"-Antwort auf Frage 31, wo er die Partei angeben soll, verneint er die Frage 30, springt er dagegen auf Frage 32.

In mündlichen und telefonischen Befragungen lassen sich mehr und komplexere Filter einsetzen, da diese Befragungen von Interviewern durchgeführt werden, die dafür geschult sind. In schriftlichen Umfragen sollte dagegen mit Filtern sparsam umgegangen werden. Sie sollten durch den Befragten leicht nachvollziehbar sein und durch ein entsprechendes Layout des Fragebogens unterstützt werden. Der ALLBUS 1990, dem wir unsere Beispiele entnommen haben, wurde mündlich durchgeführt. Computergestützte Befragungen (auch Web-Surveys) erlauben den Einsatz komplexer Filter, weil die Software die Filterführung übernimmt.

Vor allem bei einer schriftlichen Befragung ist das *Layout* des Fragebo-

Abbildung 4.4: Parteiidentifikationsfrage im ALLBUS 1990

Nr.			Weiter mit
		K.4	
30.	Viele Leute in der Bundesrepublik neigen längere Zeit einer bestimmten politischen Partei zu, obwohl sie auch ab und zu mal eine andere Partei wählen. Wie ist das bei Ihnen: Neigen Sie – ganz allgemein gesprochen – einer bestimmten Partei zu?	Ja .... ☐ 1 58	31
		Nein .... ☐ 2	32
		Verweigert .... ☐ 7	
		59/60	
31.	Sagen Sie mir bitte auch noch, welche Partei das ist? *Falls „andere Partei", nachfragen:* welche?	CDU bzw. CSU .... ☐ 01	
		SPD .... ☐ 02	
		F.D.P. .... ☐ 03	
		NPD .... ☐ 04	
		DKP .... ☐ 05	
		Die Grünen .... ☐ 06	
		Alternative Liste .... ☐ 07	
		SEW .... ☐ 08	
		Die Republikaner .... ☐ 09	
		Andere Partei, und zwar: ________	
		Verweigert .... ☐ 97	
		97/98/99	
32.	Nun zu einem ganz anderen Thema: Haben Sie schon einmal von der Krankheit AIDS gehört?	Ja .... ☐ 1 61	33
		Nein .... ☐ 2	39

gens wichtig. Eine übersichtliche Gestaltung erleichtert das Ausfüllen des Fragebogens. Bei der persönlichen und telefonischen Befragung spielt das Layout dagegen keine so große Rolle, da der Interviewer sich vor der Befragung mit dem Fragebogen (in gedruckter oder elektronischer Form) vertraut machen kann.

Ob die Befragten mit den Fragen und dem Fragebogen zurechtkommen, wird mit *Pretests* überprüft. Bei einem Pretest wird der Fragebogen und gegebenenfalls einzelne Fragen *vor* der eigentlichen Befragung an einer Stichprobe aus der Zielpopulation getestet. Pretests geben unter anderem Aufschluss über Probleme im Verständnis oder der Anordnung von

Fragen, über das Antwortverhalten, über den zeitlichen Umfang der Befragung, über Abbruchquoten etc. Für Pretests wurden verschiedene Techniken entwickelt. Zur Überprüfung des Verständnisses einzelner Fragen bieten sich kognitive Pretests an. Dabei werden die gegebenen Antworten hinterfragt, die Befragten werden aufgefordert, den Antwortprozess laut zu artikulieren oder die Frage mit eigenen Worten zu wiederholen (vgl. die Anwendung dieser Techniken bei Kurz et al. 1999). Eine quantitativ orientierte Methode zum Pretest von CATI-Interviews findet sich bei Faulbaum et al. (2003). Bevor die Umfrage ins Feld geht, muss in jedem Fall der komplette Fragebogen getestet werden. Dabei zeigt sich, ob die Filterführung und Kodierung der Variablen funktioniert, welche Fragen Nachfragen der Befragten verursachen, wie hoch die Antwortverweigerung ist, ob die Antworten der Befragten variieren usw. Die Stichprobe sollte ausreichend groß sein, um solche Probleme festzustellen.

4.1.4 Der Ablauf der Befragung

Nach der Entwicklung des Fragebogens werden zunächst die Befragten ausgewählt. Wie dies geschieht und was dabei zu beachten ist, werden wir in Kapitel 9 behandeln.

Die *Ankündigung des Interviews* soll auf den Besuch des Interviewers und das Thema der Befragung vorbereiten. Der Interviewer kann dann bei seinem Besuch noch einmal Näheres zur Fragestellung erläutern. Schriftliche und telefonische Interviews werden in der Regel nicht angekündigt, auch wenn dies möglich wäre. In einer schriftlichen Befragung wird im *Anschreiben* die Fragestellung der Untersuchung erklärt, in welchem Auftrag die Untersuchung durchgeführt wird usw. Dieses Anschreiben ist für die Teilnahmebereitschaft einer ausgewählten Person von Bedeutung. Am schwierigsten ist es sicher, einen Befragten am Telefon zur Teilnahmebereitschaft zu bewegen. Die Hemmschwelle, hier einfach aufzulegen, ist nicht besonders hoch.

Der nächste Punkt ist die *Zusicherung von Anonymität*. Grundsätzlich gilt für die Durchführung seriöser Umfragen, dass die Angaben der Befragten vertraulich behandelt werden. Adressen usw. sollten getrennt von den Antworten aufbewahrt und später vernichtet werden. Es darf keine Identifizierung, eine Verknüpfung der Daten mit anderen Informationen oder eine Weitergabe der Daten *ohne* vorherige Anonymisierung erfolgen.

Dies sollte auch dem Befragten klar gemacht werden. Ob die Anonymität der Befragung tatsächlich gewährleistet ist, kann von den Befragten allerdings nicht kontrolliert werden. Bei der mündlichen und telefonischen Umfrage, bei denen die Adresse bzw. Telefonnummer des Befragten dem Interviewer bekannt sind, sollte dem Befragten deshalb glaubhaft versichert werden, dass seine Angaben vertraulich behandelt werden. Bei der schriftlichen Befragung ist diese Zusicherung für den Befragten leichter nachvollziehbar, wenn der Fragebogen ohne Angabe eines Absenders zurückgesendet werden soll. Allerdings ist eine schriftliche Befragung nur dann anonym, wenn die Fragebögen nicht mit einem eindeutigen Code gekennzeichnet sind, der die Zuordnung zum Befragten ermöglicht. Eine Kennzeichnung der Fragebögen erfolgt häufig, um diejenigen Personen, die nach einer Weile noch keinen Fragebogen zurückgesandt haben, erneut anzuschreiben und um Beantwortung zu bitten.

Während der Befragung sollte der Befragte mit dem Interviewer alleine sein und sich ganz auf die Befragung konzentrieren können. Außerdem sollte die Befragungsperson nicht durch andere Einflüsse gestört werden. Solche *Umgebungsbedingungen* lassen sich kontrollieren, indem sie vom Interviewer erfasst werden und damit für die Auswertung der Daten verfügbar sind. Dies ist allerdings nur bei der mündlichen Befragung gut praktizierbar, während beim Telefoninterview externe Einflüsse meist nur erahnt werden können. Beim schriftlichen Interview ist eine Information über die Befragungssituation völlig unmöglich. Zu den Umgebungsbedingungen gehören auch bewusste und unbewusste Beeinflussungen durch den Interviewer. Unbewusste *Interviewereffekte* treten bei mündlichen und telefonischen Befragungen immer auf (vgl. Koch 1991; Reinecke 1991). Für das Antwortverhalten mancher Befragter spielt es eben eine Rolle, welchem Interviewer man gegenüber sitzt (jung oder alt, Mann oder Frau usw.). Auch am Telefon kann ein solcher Effekt durch die Stimme oder den Dialekt des Interviewers verursacht werden.

Das folgende Beispiel für einen Interviewereffekt basiert auf Daten des ALLBUS 1990. Dort wurde mit mehreren Statements versucht, abweichendes Verhalten zu erfassen. Der Fragetext lautete: „Bitte sagen Sie mir jeweils mit Hilfe dieser Liste, ob Sie persönlich das beschriebene Verhalten für sehr schlimm, ziemlich schlimm, weniger schlimm oder für überhaupt nicht schlimm halten.“ Eines der Statements auf der Liste lautete: „Ein Mann zwingt seine Ehefrau zum Geschlechtsverkehr.“ Wie man der nachfolgenden Tabelle 4.3 entnehmen kann, unterscheidet sich das Antwort-

verhalten der Männer danach, ob sie von einem Mann oder einer Frau interviewt werden. Während das Geschlecht des Interviewers das Antwortverhalten der weiblichen Befragten nicht beeinflusst, zeigt sich bei den männlichen Befragten, dass diese eine Vergewaltigung in der Ehe für weniger schlimm halten, wenn sie von Männern interviewt werden. So sagen zwar rund 70 % der Männer, die von Frauen interviewt werden, das beschriebene Verhalten sei „sehr schlimm“, aber nur 62,5 % der Männer, die von Männern interviewt werden. (Es handelt sich um einen Interaktionseffekt, wie wir aus Kapitel 2.3 wissen.)

Tabelle 4.3: Interviewereffekt bei der mündlichen Befragung (Spaltenprozente)

	Männliche Befragte		Weibliche Befragte	
Antwort	Männl. Int.	Weibl. Int.	Männl. Int.	Weibl. Int.
Sehr schlimm	62,5 %	69,9 %	71,1 %	70,7 %
Ziemlich schlimm	29,1 %	23,0 %	24,4 %	25,6 %
Weniger schlimm	6,7 %	5,6 %	3,9 %	2,9 %
Gar nicht schlimm	1,8 %	1,5 %	0,6 %	0,8 %
Befragte	506	196	492	242

Quelle: ALLBUS 1990

Der Interviewer kann auch bewusst eine Befragung beeinflussen oder sogar insgesamt fälschen (vgl. Dorroch 1994; Koch 1995; Schnell 1991). Solche Formen von Betrug gibt es immer wieder, vor allem bei mündlichen Befragungen, wo der Interviewer das Interview weitgehend selbst in der Hand hat. Er kann zum Beispiel absichtlich eine falsche Filterführung anwenden, um das Interview abzukürzen. Oder er füllt gleich den ganzen Fragebogen selbst aus. Solchen Interviewerfälschungen kommt man auf die Spur, indem man die für das Interview ausgewählten Personen anruft und nachfragt, ob auch tatsächlich ein Interview stattgefunden hat. Mit dieser Nachfrage kann man allerdings nicht kontrollieren, ob ein Interviewer lediglich einen Teil des Interviews ordnungsgemäß durchgeführt hat und den „Rest“ des Fragebogens selbst ergänzt hat. Telefonische Umfragen bieten dagegen eine umfassende Kontrollmöglichkeit der Interviewer, wenn sie in einem Telefonstudio durchgeführt werden, wo ein „Supervisor“ die Durchführung der Interviews überwachen kann. Auch computergestützte persönliche Befragungen sind schwerer zu fälschen.

4.2 Beobachtung

Die Beobachtung (vgl. Grümer 1974; Friedrichs und Lüdtke 1977) wird in den Sozialwissenschaften eher stiefmütterlich behandelt, obwohl sie für einen Teil der Fragestellungen durchaus sinnvoll verwendet werden könnte. Als eines der wenigen politikwissenschaftlichen Beispiele kann hier die von Raschke (1991) durchgeführte Beobachtung einer Bundesdelegiertenversammlung der GRÜNEN genannt werden.

Wissenschaftliche Beobachtungen unterscheiden sich von Beobachtungen im Alltag vor allem durch systematisches und an Hypothesen orientiertes Vorgehen. Ziel der Beobachtung ist das „Erfassen von Ablauf und Bedeutung einzelner Handlungen und Handlungszusammenhänge“ (Kromrey 2006, 346). Im Vergleich zur Befragung, wo über Verhalten lediglich Aussagen gewonnen werden, richtet sich die Beobachtung direkt auf das *Verhalten* der Subjekte.

Eine besondere Schwierigkeit bei der Durchführung einer Beobachtung besteht darin, alle Ereignisse einer sich ständig verändernden Situation wahrzunehmen. Welche Ereignisse registriert werden, hängt stark von der selektiven Wahrnehmung des Beobachters ab. Aus diesem Grund ist die Beobachtung ein relativ subjektives Verfahren. Eine Kontrolle, ob der Beobachter die für die Untersuchung bedeutsamen Ereignisse erfasst hat, ist normalerweise nicht möglich, es sei denn, die Beobachtung wird mit technischen Hilfsmitteln, also z. B. Film- oder Videokamera, aufgezeichnet. Auch Filmaufnahmen sind jedoch selektiv und liefern keine ‚objektiven‘ Informationen. Wie bei der Befragung der Interviewer, stellt bei der Beobachtung der Beobachter also eine mögliche Fehlerquelle dar, die zu einer Verzerrung der Resultate führen kann. Die einzige Möglichkeit, solche Fehler zu minimieren, besteht in der Schulung der Beobachter.

Nach Friedrichs (1990) lassen sich Beobachtungen danach unterscheiden, ob sie *verdeckt* oder *offen*, *teilnehmend* oder *nicht teilnehmend*, *strukturiert* oder *unstrukturiert*, in *künstlichen* oder *natürlichen* Situationen stattfinden.

- Bei einer *verdeckten* Beobachtung nehmen die Beobachteten den Beobachter nicht wahr, während bei einer *offenen* Beobachtung – z. B. der oben erwähnten Analyse des Verhaltens der Parteitagsdelegierten – der Beobachter als solcher auch auftritt.

Eine verdeckte Beobachtung wird dann notwendig, wenn ansonsten überhaupt keine Beobachtung stattfinden könnte. Interessiert man sich für interne Strukturen einer verbotenen politischen Vereinigung, so ist eine offene Beobachtung wahrscheinlich nicht möglich. Ethisch bedenklich sind verdeckte Beobachtungen immer, wenn man davon ausgehen kann, dass die Beobachteten Einwände gegen die Beobachtung haben. In einem solchen Fall muss man schon gewichtige Gründe ins Feld führen, um die Beobachtung zu rechtfertigen.
Verdeckte Beobachtungen besitzen gegenüber offenen Beobachtungsformen allerdings den Vorteil, dass sie die zu beobachtenden Ereignisse nicht beeinflussen, also ein nicht reaktives Messverfahren darstellen.

- Bei einer *teilnehmenden* Beobachtung ist der Beobachter aktiv ins Geschehen einbezogen, während er bei einer nicht teilnehmenden Beobachtung auf die Beobachtung beschränkt bleibt. Bei der *nicht teilnehmenden* Beobachtung eines Parteitages z. B. kommt dem Beobachter ein reiner Beobachtungsstatus zu. Eine teilnehmende Beobachtung läge dann vor, wenn der Beobachter gleichzeitig Parteitagsdelegierter wäre.
  Bei einer teilnehmenden Beobachtung werden die Ereignisse, die beobachtet werden sollen, durch das *Handeln* des Beobachters beeinflusst. Zudem besteht die Gefahr, dass der Beobachter durch seine aktive Rolle die Distanz zum sozialen Geschehen verliert und dadurch die Perspektive der zu Beobachtenden – z. B. der Parteitagsdelegierten – annimmt. Außerdem könnte der Beobachter auch einfach überfordert sein, denn schließlich verlangt eine teilnehmende Beobachtung nicht nur die sorgfältige Wahrnehmung der Situation, sondern gleichzeitig auch noch aktives Handeln.
- Einer *strukturierten* Beobachtung liegt ein zuvor entwickeltes detailliertes Kategorienschema zugrunde, in das die Beobachtungen eingetragen werden. Bei unstrukturierten Beobachtungen ist die Beobachtung in einem erheblich geringeren Umfang durch vorherige Festlegungen geprägt. *Unstrukturierte* Beobachtungen sind vor allem bei geringen Kenntnissen über die zu beobachtende Situation sinnvoll. In diesem Fall hat die Beobachtung zunächst einmal explorativen Charakter.
- *Künstlich* ist eine Beobachtung dann, wenn die Beobachtung in einer Laborsituation stattfindet, d. h. die Umgebungsbedingungen gezielt hergestellt und verändert werden können. Solche „Experimente" haben den entscheidenden Vorteil, dass man Kausalität nachprüfen kann.

Durch die Kontrollgruppen und gesteuerten Versuchsbedingungen ist die interne Validität der Laborbeobachtung sehr hoch. Wie bereits in Kapitel 2 deutlich wurde, sind jedoch nicht alle Fragestellungen für Experimente geeignet. Vor allem die mangelnde Übertragbarkeit auf natürliche Umgebungen spricht gegen künstliche Beobachtungen, d. h. sie weisen nur eine geringe externe Validität auf.

Diese Unterscheidungskriterien lassen sich nun auf vielfältige Art und Weise zu unterschiedlichen Typen von Beobachtungen kombinieren. In der Praxis sind allerdings nur einige wenige Kombinationen von Interesse. Für Politikwissenschaftler kommen vor allem offene, systematische Beobachtungen in natürlichen Umgebungen in Frage. Möchte man Entscheidungsstrukturen auf Parteitagen der unterschiedlichen Parteien untersuchen, so ist es beispielsweise nicht notwendig, verdeckt zu arbeiten. Im folgenden werden wir uns – wie schon bei der Befragung – auf eine *strukturierte* Erhebungsform beschränken.

4.2.1 Kategorienentwicklung

Während die Messinstrumente bei einer Befragung aus einer oder mehreren Fragen bestehen, stellt die *Beobachtungskategorie* das Messinstrument bei der Beobachtung dar. Als Messinstrument für die Beteiligung von Delegierten und Vorstandsmitgliedern an einem Parteitag könnten z. B. die Beobachtungskategorien „Anzahl der Redebeiträge von Delegierten" und „Anzahl der Redebeiträge von Vorstandsmitgliedern" herangezogen werden. Die Merkmalsausprägungen sind bei diesen beiden Variablen die Häufigkeit. Zudem könnte die „Redelänge" oder die „Reaktion der Delegierten" erfasst werden. Die Merkmalsausprägungen könnten hier z. B. die Länge in Minuten bzw. die Klassifizierungen „zurückhaltend", „enthusiastisch" usw. sein. Für die Beobachtungskategorien gilt dasselbe wie für die Antwortvorgaben bei einer Befragung: Sie müssen sich gegenseitig ausschließen und vollständig sein.

Nehmen wir an, wir interessieren uns dafür, ob sich die Beteiligung von Männern und Frauen im Studierendenparlament (im folgenden StuPa) unterscheidet. Unter „Beteiligung" können wir die Anwesenheit, Redebeiträge und Zwischenrufe in StuPa-Sitzungen verstehen. Damit haben wir also drei abhängige Variablen, die als Indikator für politische Beteiligung im StuPa dienen. Als unabhängige Variable müssen wir auf jeden Fall das

Geschlecht heranziehen. Auch hier ist natürlich wichtig, dass keine Beobachtungskategorie (dies entspricht einer Frage im Fragebogen) vergessen wird. Es könnte ja sein, dass die Anwesenheit nicht nur vom Geschlecht, sondern auch von der politischen Einordnung abhängt. Wir könnten z. B. vermuten, dass die Mitglieder der Opposition im StuPa in einem geringeren Umfang ihr Mandat wahrnehmen als die regierenden Parteien. Wir müssen also bei jeder und jedem Anwesenden vermerken, ob es sich um eine Frau oder einen Mann, ein Mitglied der Opposition oder der Koalition handelt. Auch der zweite Indikator für politische Beteiligung, die Redebeiträge, könnten auf diese Art und Weise operationalisiert werden. Zusätzlich könnten wir hier noch die Länge für relevant erachten. Vielleicht gibt es zwar nicht mehr Redebeiträge von Männern als von Frauen, Männer und Frauen könnten sich jedoch in der Dauer ihrer Redebeiträge unterscheiden. Ebenso wie die Anwesenheit und die Redebeiträge könnten wir noch die Zwischenrufe nach dem Geschlecht notieren.

Bereits die wenigen hier verwendeten Kategorien werden die Aufmerksamkeit des Beobachters in vollem Umfang in Anspruch nehmen. Zudem wird vorausgesetzt, dass er sich mit der Situation relativ gut auskennt. Um die politische Richtung zu notieren, muss der Beobachter alle im StuPa anwesenden Mitglieder einordnen können. Sind im StuPa viele Gruppierungen vertreten, so wird diese Einordnung schon ziemlich schwierig. Die gleichzeitige Erfassung von Redebeiträgen und Zwischenrufen ist für einen einzigen Beobachter alleine wohl nicht mehr durchführbar.

Schon an diesem kleinen Beispiel wird deutlich, warum ein Beobachtungsschema in der Regel weniger Kategorien enthält als ein Fragebogen Fragen beinhaltet: Die Aufmerksamkeit des Beobachters kann sich immer nur auf einige wenige Merkmale richten.

4.2.2 Beobachtungsschema

Alle Beobachtungskategorien werden in einem *Beobachtungsschema* bzw. Beobachtungsprotokoll zusammengefasst. Dem Beobachtungsprotokoll entspricht bei der Befragung der Fragebogen. Anhand dieses Beobachtungsprotokolls wird die Beobachtung durchgeführt.

Die Zusammenfassung der Beobachtungskategorien in ein Beobachtungsschema sollte so erfolgen, dass der Beobachter möglichst schnell seine Eintragungen vornehmen kann. Ob das Beobachtungsschema praktikabel ist

und alle relevanten Kategorien erfasst wurden, kann in einer Testbeobachtung geprüft werden. Ein fiktives Beobachtungsschema für die Anwesenheit und die Redebeiträge in StuPa-Sitzungen ist in Abbildung 4.5 zu sehen: Zunächst trägt der Beobachter einen Namen, das Datum der Sitzung und die Uhrzeit ein. Danach folgt eine Tabelle, in der für jedes anwesende StuPa-Mitglied eine Zeile vorgesehen sein muss, in die das Geschlecht und die politische Zugehörigkeit einzutragen ist (maximal sind hier also so viele Zeilen wie Mandatsträger notwendig). Auf einem/mehreren neuen Blättern können dann die Redebeiträge festgehalten werden. Bei der Interpretation der Daten muss man beachten, dass das eine Mal Merkmale von Personen, das andere Mal Merkmale von Redebeiträgen erhoben werden.

4.2.3 Ablauf einer Beobachtung

Zunächst müssen die Situationen festgelegt werden, die beobachtet werden sollen. Wollen wir die Hypothese untersuchen, dass sich die Entscheidungsstrukturen auf den Parteitagen zwischen den einzelnen Parteien unterscheiden, dann sind die relevanten Situationen Parteitage. Zudem müssen wir den Untersuchungszeitraum festlegen und eine räumliche Abgrenzung vornehmen. Im genannten Beispiel könnten dies die Parteitage innerhalb eines Jahres in der Bundesrepublik Deutschland sein.

Ebenso wichtig wie die Schulung von Interviewern ist die Schulung der Beobachter. Aufgrund der Selektivität der Wahrnehmung werden unterschiedliche Beobachter zu unterschiedlichen Ergebnissen gelangen. Das Ausmaß, in dem verschiedene Beobachter dieselben Dinge wahrnehmen, wird *Inter-Coder-Reliabilität* genannt. Durch genaue Anweisungen und Beispiele kann die Inter-Coder-Reliabilität erhöht werden. Daneben kann es im Verlauf einer Beobachtung dazu kommen, dass der Beobachter – z. B. durch einen Lernprozess – dieselben Ereignisse unterschiedlich einordnet. Die Stabilität der Einordnung derselben Beobachtungsinhalte durch einen Beobachter wird *Intra-Coder-Reliabilität* genannt. Das Ausmaß der Inter-Coder-Reliabilität und der Intra-Coder-Reliabilität ist ein Maß für die Qualität der Beobachtung.

Möglicherweise ist es – wie erwähnt – sinnvoll, das Beobachtungsschema vorab auf Praktikabilität zu testen. Bei manchen offenen Beobachtungen wie den Parteitagen oder den StuPa-Sitzungen bietet es sich an, die Beobachtung vorher anzukündigen.

Abbildung 4.5: Fiktives Beobachtungsprotokoll einer StuPa-Sitzung

ALLGEMEINE ANGABEN

Beobachter:
Datum:
Uhrzeit (Beginn und Ende):

ANWESENDE

	Geschlecht		*Politische Einordnung*	
Nr.	weiblich	männlich	Koalition	Opposition
1				
2				
3				
4				
5				
6				
7				
8				
9				
u.s.w.	⋮	⋮	⋮	⋮

REDEBEITRÄGE

	Geschlecht		*Politische Einordnung*		*Länge*
Nr.	weiblich	männlich	Koalition	Opposition	Minuten
1					
2					
3					
4					
5					
6					
7					
8					
9					
u.s.w.	⋮	⋮	⋮	⋮	⋮

4.3 Inhaltsanalyse

Bei der Inhaltsanalyse gewinnt man Informationen über die soziale Wirklichkeit durch die Analyse von „Inhalten“. Inhalte können *Texte, Filme, Bilder* o. Ä. sein. Das Ziel der Inhaltsanalyse kann man auf den Nenner bringen: „Wer sagt was zu wem, wie, warum und mit welchem Effekt?“ (Friedrichs 1990, 319). Die Anwendung der Inhaltsanalyse (vgl. Früh 2007; Merten 1995; Weber 1990) erfolgt vor allem im Bereich der Massenkommunikation und hat deshalb innerhalb der Publizistik einen hohen Stellenwert. In der Politikwissenschaft bieten sich zahlreiche Verwendungs-

möglichkeiten wie die Analyse von Parteiprogrammen, politischen Reden, Flugblättern und Wahlplakaten an (vgl. Rucht et al. 1995).

Der wesentliche *Vorteil der Inhaltsanalyse* besteht darin, dass die Inhalte nicht zum Zweck der Untersuchung geschaffen wurden. Zeitungsartikel werden für die Leser der Zeitung geschrieben, nicht für den Wissenschaftler, der die Artikel nach bestimmten Kriterien auswertet. Anders ausgedrückt: Die Inhalte werden nicht vom Erhebungsinstrument bestimmt oder beeinflusst. Aus diesem Grund ist die Inhaltsanalyse ein *nicht reaktives Messverfahren*, d. h. die Ergebnisse der Untersuchung hängen nicht von der Methode ab. Wie bei der Befragung und der Beobachtung unterscheidet sich die wissenschaftliche von der alltäglichen Inhaltsanalyse durch die systematisierte und objektivierte Vorgehensweise. Auch Inhaltsanalysen können nach dem Grad der Strukturierung unterschieden werden; wir werden lediglich auf die quantitative Inhaltsanalyse eingehen (vgl. zur qualitativen Inhaltsanalyse Mayring 2007).

Zunächst muss festgelegt werden, was überhaupt analysiert werden soll. Wollen wir z. B. die inhaltlichen Schwerpunkte der Parteien nach Politikbereichen untersuchen, dann könnten die Parteiprogramme die Textgrundlage sein. Für diese Texte muß der Merkmalsträger, d. h. die *Zähleinheit* bestimmt werden. Als Zähleinheit könnten wir z. B. die Sätze der Parteiprogramme verwenden. Ebenso gut könnten aber auch kleinere Einheiten (z. B. einzelne Wörter) oder größere Einheiten (z. B. Absätze oder einzelne Abschnitte des Parteiprogramms) ausgewählt werden. Den Zähleinheiten entsprechen bei der Beobachtung die Situationen, bei der Befragung die Befragten.

Der schwierigste Teil einer Inhaltsanalyse besteht darin, Kategorien zu entwickeln, die die theoretischen Begriffe messen. An das *Kategorienschema* sind hier die gleichen Anforderungen zu stellen wie bei der Beobachtung: Sie müssen sich gegenseitig ausschließen und vollständig sein. Kategorien für den Untersuchungsgegenstand „Politikbereiche“ könnten z. B. „Innenpolitik“, „Außenpolitik“, „Wirtschaftspolitik“ usw. sein. Auch hier kann eine Kategorie „Sonstiges“ für nicht explizit genannte Kategorien sinnvoll sein. Das Kategorienschema ist in der Regel natürlich nicht so einfach wie im vorgestellten Beispiel. Aus diesem Grund ist es sinnvoll, erst einmal einen Pretest an einer geringen Anzahl von Analyseeinheiten durchzuführen und die Kategorien unter Umständen abzuändern.

Die Kodierung ist die Zuordnung der Zähleinheiten zu den Kategorien. Die Inhaltsanalyse ist abgeschlossen, wenn alle Analyseeinheiten kodiert sind. Im vorliegenden Beispiel könnte man z. B. die Analyseeinheiten der einzelnen Parteiprogramme zusammenfassen. Für jedes Programm könnten wir dann sagen, wie häufig auf die Wirtschaftspolitik, Sozialpolitik usw. Bezug genommen wurde. Wenn man davon ausgeht, dass aus der Häufigkeit einer Kategorie auf die Bedeutung eines Politikbereiches für eine Partei geschlossen werden kann, könnte das Ergebnis einer Inhaltsanalyse sein, dass das SPD-Programm stärker sozialpolitisch bestimmt wird als das CDU-Programm, der Schwerpunkt des FDP-Programms dagegen auf der Wirtschaftspolitik liegt usw.

Einfache Inhaltsanalysen begnügen sich damit, die Häufigkeit des Vorkommens der einzelnen Kategorien auszuwerten („Frequenzanalyse"). Weitergehende inhaltsanalytische Ansätze berücksichtigen zudem Bewertungen („Bewertungsanalyse") oder Zusammenhänge zwischen Kategorien („Kontingenzanalyse"), d. h. wie häufig tauchen die einzelnen Kategorien im Zusammenhang mit anderen Kategorien auf. Bei einer Bewertungsanalyse der Parteiprogrammatik würde man also nicht nur eine Aussage zur Gewerbesteuer als Aussage zum Politikbereich Wirtschaft kennzeichnen, sondern berücksichtigen, ob diese negativ oder positiv bewertet wird.

Die Zuverlässigkeit der Inhaltsanalyse beinhaltet auch hier wieder zwei Aspekte: Die *Inter-Coder-Reliabilität* ist hoch, wenn verschiedene Vercoder dieselben Analyseeinheiten in dieselben Kategorien einordnen. Die *Intra-Coder-Reliabilität* ist hoch, wenn die Zuordnung einer Analyseeinheit zu einer Kategorie durch einen einzigen Vercoder stabil ist. Die Inter-Coder-Reliabilität lässt sich hier sehr einfach überprüfen, indem man unterschiedlichen Codierern dieselben Texte vorlegt.

Stellt man fest, dass das Kategorienschema nicht angemessen war, also z. B. wesentliche Dimensionen fehlten, dann kann man die Inhalte problemlos erneut auswerten. Dies ist ein wesentlicher Vorteil der Inhaltsanalyse gegenüber der Beobachtung.

Aufgaben zu Erhebungsmethoden

1. Studierende der Politikwissenschaft sollen zu ihrem Studium befragt werden. Was ist an den einzelnen Fragen „faul"? Wie können diese besser formuliert werden?
   a) Wie hoch schätzen Sie die durchschnittlichen Kosten eines Hochschulstudiums ein?
      ____________ Euro
   b) Sind Sie für eine Straffung des Studiums und die Einführung von Studiengebühren für Langzeitstudierende?
      Ja .... ☐
      Nein .... ☐
   c) Wie häufig essen Sie in der Mensa?
      Täglich .... ☐
      Zwei- bis dreimal wöchentlich .... ☐
      Mehr als einmal wöchentlich .... ☐
      Seltener .... ☐
      Nie .... ☐
      Weiß nicht .... ☐
   d) Welche Gründe waren ausschlaggebend für die Wahl ihres Studienfachs? (Mehrfachnennungen möglich!)
      Gute Berufsperspektiven .... ☐
      Der Studienort .... ☐
      Selbstverwirklichung .... ☐
      Ich möchte in die Politik .... ☐
   e) Sind Sie nicht der Meinung, dass der AStA nicht die Interessen der Studierenden vertritt?
      Ja .... ☐
      Weiß nicht .... ☐
      Nein .... ☐
   f) Halten Sie es für angemessen, dass Studierende, die von der Nutzung öffentlicher Verkehrsmittel, die im Rahmen der biannual anfallenden Semesterbeiträge bereits vorfinanziert wurde, absehen, eine reduzierte Studiengebühr entrichten?
      Ja .... ☐
      Nein .... ☐
      Weiß nicht .... ☐
   g) In welchem Fachsemester studieren Sie? ____________

2. Wann bieten sich geschlossene Fragen an, wann offene?

5 Tabellen und Graphiken

Sobald die Datenerhebung abgeschlossen ist, kann mit der Auswertung begonnen werden. In der Regel werden die Daten zunächst in maschinenlesbare Form gebracht, um sie dann mit Hilfe eines Statistik-Programms wie z. B. SPSS, SAS oder Stata auswerten zu können. Dies ist jedoch nicht immer notwendig. Einfache Analysen können – wenn auch mit einem erheblich höheren Zeitaufwand – mit der Hand bzw. dem Taschenrechner durchgeführt werden.

Man kann sich leicht vorstellen, dass schon in einer Umfrage mit wenigen Befragten der Überblick über die Antworten zur „Wahlsonntagsfrage“, die die Wahl*absicht* der Befragten misst, ohne eine Zusammenfassung der Nennungen für die verschiedenen Parteien verloren geht. In der Regel beginnt deshalb jede Analyse mit einer Häufigkeitsauszählung der interessierenden Merkmalsausprägungen, die dann tabellarisch oder graphisch dargestellt werden. Untersucht man die Verteilung eines einzigen Merkmals, dann spricht man von *univariater* Analyse. Betrachtet man dagegen die gemeinsame Verteilung von zwei oder mehr Merkmalen, dann spricht man von *bivariater* bzw. *multivariater* Analyse.

5.1 Tabellen

5.1.1 Tabellarische Darstellung *eines* Merkmals

Zunächst will man in der Regel wissen, wie stark die einzelnen Ausprägungen *einer* Variable besetzt sind. Wie viele Befragte geben z. B. an, am nächsten Sonntag die CDU, SPD usw. zu wählen? Um dies herauszufinden, führt man eine *Häufigkeitsauszählung* der einzelnen Kategorien durch. Als Resultat erhält man eine *Häufigkeitsverteilung.*

Üblicherweise wird ein Merkmal bzw. eine Variable mit einem großen lateinischen Buchstaben bezeichnet. Beispielsweise soll es um die Wahlabsicht gehen. Diese wird dann als X bezeichnet. Dieses Merkmal X kann ganz bestimmte Merkmalsausprägungen annehmen („CDU/CSU“, „SPD“ etc.). Die Merkmalsausprägungen werden als x_k bezeichnet, wobei der Index k

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_5,

eine fortlaufende Nummerierung der Merkmalsausprägungen (Kategorien) meint und dementsprechend von 1 bis zur maximalen Merkmalsausprägung m läuft. k wird deshalb als Laufvariable oder Laufindex bezeichnet und tief gestellt (vgl. Tabelle 5.1).

Tabelle 5.1: Notation bei Häufigkeitsauszählungen

Kategorie (Merkmalsausprägung)	Bezeichnung
CDU/CSU	x_1
SPD	x_2
FDP	x_3
NPD	x_4
B' 90/GRÜNE	x_5
REP	x_6
Andere Partei	x_7
Wähle nicht	x_8
Verweigert	x_9
Weiß nicht	x_{10}
Keine Angabe	x_{11}

Absolute Häufigkeiten geben die **Anzahl** der Merkmalsträger wieder, die eine bestimmte Merkmalsausprägung aufweisen. Absolute Häufigkeiten einer Merkmalsausprägung sind ohne eine Berücksichtigung der Gesamtzahl der Merkmalsträger nicht interpretierbar. Wenn 100 Befragte CDU wählen wollen, sagt das gar nichts über die Chancen der CDU aus, wenn man nicht weiß, wie viele Befragte *insgesamt* eine Wahlabsicht geäußert haben. Man muss also wissen, welchen **Anteil** eine absolute Häufigkeit an allen Häufigkeiten hat. **Relative Häufigkeiten** (*Anteil der Merkmalsträger* in einer bestimmten Kategorie) oder **Prozentwerte** (relative Häufigkeit $\times$ 100) werden berechnet, indem die absolute Häufigkeit einer Kategorie ins Verhältnis zur Gesamtzahl der Merkmalsträger n gesetzt wird. Merkmalsträger werden auch als Fälle bezeichnet.

Absolute Häufigkeiten werden als f_{x_k} bezeichnet (f steht für engl. *frequency* = Häufigkeit). Für relative Häufigkeiten gibt es keine eigene Notation, Prozentwerte werden durch das nachgestellte Prozentzeichen (%) kenntlich gemacht.

$$\text{Absolute Häufigkeit} = f_{x_k}$$

$$\text{Relative Häufigkeit} = \frac{f_{x_k}}{\text{Gesamtzahl der Merkmalsträger}} = \frac{f_{x_k}}{n}$$

$$\text{Prozentwert} = \text{Relative Häufigkeit} \times 100$$

Absolute Häufigkeiten können Werte zwischen 0 und $+\infty$ annehmen. Daraus folgt, dass relative Häufigkeiten immer einen Bruch zwischen zwei positiven Zahlen darstellen, wobei der Nenner **immer** größer oder gleich dem Zähler ist, da die absolute Häufigkeit *einer* einzelnen Merkmalsausprägung nicht größer als die Gesamtzahl der Merkmalsträger sein kann. Relative Häufigkeiten können deshalb nur Werte zwischen 0 und 1 annehmen. Da ein Prozentwert einfach die Multiplikation einer relativen Häufigkeit mit 100 darstellt, können Prozentwerte nur im Bereich von 0 % bis 100 % liegen.

In Tabelle 5.2 auf der gegenüberliegenden Seite ist eine Häufigkeitsauszählung der „Wahlsonntagsfrage“ aus dem ALLBUS 1994 dargestellt.[1] 692 Befragte gaben an, CDU/CSU zu wählen, wenn am nächsten Sonntag Wahlen stattfänden. Ohne die Gesamtzahl der Befragten (2.298) zu berücksichtigen, ist dieser Wert nicht sehr aussagekräftig. Aus diesem Grund sind in den beiden darauf folgenden Spalten die relativen Häufigkeiten und die Prozentwerte angegeben. Demnach würden sich 30,1 % der Befragten für die CDU/CSU entscheiden. Da relative Häufigkeiten und Prozentwerte dieselben Informationen liefern, gibt man in Tabellen nur einen der beiden Werte – in der Regel die Prozentwerte – an.

Prozentzahlen sind natürlich nur dann sinnvoll interpretierbar, wenn die **Größe der Prozentuierungsbasis** bekannt ist. Es macht nämlich für die Bedeutung des Wertes einen Unterschied, ob 75 % der Befragten, die ein bestimmtes Waschmittel favorisieren, 3 von insgesamt 4 befragten Personen oder 3.000 von 4.000 sind. Im ersten Fall würde der Wechsel einer Person in das „andere Lager“ gleich das Verhältnis auf 100 % bzw. 50 % verändern. Im letzten Fall würde sich der Wechsel einer Person lediglich in einer Verschiebung auf 75,025 % bzw. 74,975 % ausdrücken.

1 In der Stichprobe sind Ostdeutsche überrepräsentiert (ca. 32 % Ost- und 68 % Westdeutsche). Da es vorläufig um die Beschreibung von Stichprobendaten geht, haben wir auf eine personenbezogene Ost-West-Gewichtung verzichtet.

Tabelle 5.2: Häufigkeitsauszählung der Wahlabsicht im ALLBUS 1994

Wahlabsicht x_k	absolute Häufigkeit f_{x_k}	relative Häufigkeit	Prozente %
CDU/CSU	692	0,301	30,1
SPD	856	0,372	37,2
FDP	200	0,087	8,7
Bündnis 90/Grüne	316	0,138	13,8
Republikaner	72	0,031	3,1
PDS	120	0,052	5,2
Andere Partei	42	0,018	1,8
Summe	2298	1,000	100,0

Genauso bedeutend wie die Größe der Prozentuierungbasis ist die **Art der Prozentuierungsbasis**. Als Beispiel soll wiederum die Wahlabsicht dienen. Dazu betrachten wir Tabelle 5.3 auf der folgenden Seite. In der Spalte „abs. H." werden die *absoluten Häufigkeiten*, in der Spalte „%" die Prozentwerte wiedergegeben. Unter der Spalte „Alle" werden die absoluten Häufigkeiten aller Kategorien aufgelistet. Von allen 3.450 Befragten in dieser Umfrage gaben 692 an, CDU/CSU wählen zu wollen, 856 die SPD usw. 570 Befragte wussten noch nicht, was sie wählen wollen.

In der dritten und vierten Spalte der Tabelle wurden nur die Befragten berücksichtigt, die eine Partei angegeben haben; bei einer Wahl wären dies die gültigen Stimmen. Die Prozentwerte der vierten Spalte geben daher die *Anteile der Parteien an den „gültigen Stimmen"* wieder. In der fünften und sechsten Spalte der Tabelle wurden dagegen die Antworten aller wahlberechtigten Befragten betrachtet. Die Prozentwerte der sechsten Spalte können daher als *Anteil der Parteien an den Wahlberechtigten* bezeichnet werden. Wie man leicht feststellen kann, unterscheiden sich die Prozentangaben der vierten und sechsten Spalte beträchtlich.

Bei Prozentangaben ist außerdem zu beachten, ob es sich bei diesen tatsächlich um relative Häufigkeiten (× 100) oder um Angaben der Größen*veränderung* handelt. Angaben der Größenveränderung lassen sich nämlich als Steigerungsrate oder als Differenz zweier Prozentangaben ausdrücken. So haben z. B. Bündnis 90/Grüne bei der Bundestagswahl 1990

Tabelle 5.3: Häufigkeitsauszählung der Wahlabsicht mit unterschiedlicher Prozentuierungsbasis

		Art der Prozentuierungsbasis			
	Alle	Abg. gültige Stimmen		Wahlberechtigte	
Wahlabsicht	abs. H.	abs. H.	%	abs. H.	%
CDU-CSU	692	692	30,1	692	21,0
SPD	856	856	37,2	856	26,0
FDP	200	200	8,7	200	6,1
Bündnis 90/Grüne	316	316	13,8	316	9,6
Republikaner	72	72	3,1	72	2,2
PDS	120	120	5,2	120	3,6
Andere Partei	42	42	1,8	42	1,3
Würde nicht wählen	245	–	–	245	7,4
Verweigert	145	–	–	145	4,4
Weiß nicht	570	–	–	570	17,3
Keine Angabe	36	–	–	36	1,1
Nicht wahlberechtigt	156	–	–	–	–
Summe	3450	2298	100,0	3294	100,0

zusammen einen Anteil von 5,05 % der Zweitstimmen erzielt.[2] Bei der Bundestagswahl 1994 lag der Stimmanteil bei 7,3 % der Zweitstimmen. Dies kann man einmal als Steigerung von 2,25 *Prozentpunkten* (7,3 - 5,05) ausdrücken oder als Steigerungs*rate* von 44,6 *Prozent*, um die der Anteil 1994 höher ausgefallen ist als 1990 [(7,3 - 5,05)/5,05 × 100].

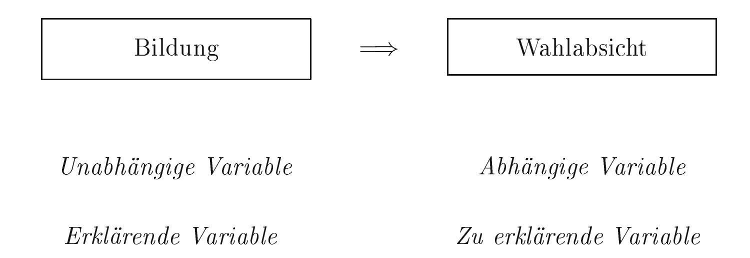
5.1.2 Kreuztabellen

Mit Hilfe von Kreuztabellen wird die *gemeinsame Verteilung von zwei Merkmalen* abgebildet. Da mit Hilfe von Kreuztabellen ein Zusammenhang zwischen Merkmalen festgestellt werden kann, spricht man auch von *Kontingenztabellen* oder *Kontingenztafeln*.

2 Die Grünen und das Bündnis ’90 traten zur Wahl 1990 noch nicht als vereinte Partei an. Da die 5 %-Klausel auf die Wahlgebiete West und Ost getrennt angewandt wurde, konnte das Bündnis 90 mit 6,05 % im Wahlgebiet Ost die 5 %-Hürde überwinden, die Grünen scheiterten mit 4,75 % im Wahlgebiet West jedoch an der 5 %-Klausel.

Unabhängige und abhängige Variablen

Bei bi- und multivariaten Analysen wird häufig zwischen *abhängigen* und *unabhängigen* Variablen unterschieden. Als **abhängig** werden die Variablen bezeichnet, die erklärt werden sollen, weshalb diese auch *zu erklärende Variablen* genannt werden. Als **unabhängig** werden die Variablen betrachtet, die (vermutlich) einen Einfluss auf die abhängige Variable ausüben. Die unabhängigen Variablen werden auch *erklärende Variablen* genannt. Zum Beispiel könnte die Wahlabsicht als abhängige und die Bildung als unabhängige Variable betrachtet werden. Wir vermuten also, dass die Wahlabsicht vom Bildungsniveau beeinflusst wird. Zur Kennzeichnung der Richtung des Zusammenhangs wurde im folgenden Schaubild ein Pfeil verwendet, dessen Spitze auf die abhängige Variable gerichtet ist.



Mit der Unterscheidung von unabhängiger und abhängiger Variable wird ein *kausaler* Einfluss der unabhängigen auf die abhängige Variable *unterstellt*; im Beispiel also, dass verschiedene Bildungsabschlüsse ein unterschiedliches Wahlverhalten verursachen. In Ex-Post-Facto-Designs soll durch die Kontrolle von Drittvariablen sichergestellt werden, dass ein statistischer Zusammenhang nicht für eine kausale Beziehung gehalten wird (vgl. Kapitel 2.3). Deshalb muss man sich vor der Durchführung einer Untersuchung, spätestens jedoch vor der Analyse bi- oder multivariater Zusammenhänge, Gedanken um mögliche „dritte" Einflussfaktoren auf die zu erklärende Variable machen.

Ob eine bestimmte Variable als abhängig oder unabhängig betrachtet wird, kann von Untersuchung zu Untersuchung und selbst innerhalb einer Untersuchung wechseln. Der eine Forscher möchte die Wahlabsicht durch die Bildung erklären (also: *Bildung* → *Wahlabsicht*), ein anderer die Bildung durch den sozialen Status des Elternhauses (also: *sozialer Status der*

Eltern → *Bildung*). Die Bestimmung der abhängigen Variable resultiert aus dem Untersuchungsinteresse, die Bestimmung der unabhängigen Variable(n) aus der zugrunde gelegten Theorie bzw. den Hypothesen.

In einer Kreuztabelle werden alle Kombinationen der Merkmalsausprägungen zweier Variablen ausgezählt. Es entstehen so Zeilen und Spalten einer Tabelle. In den deutschen Sozialwissenschaften wird die unabhängige Variable normalerweise in den Spalten und die abhängige Variable in den Zeilen abgetragen. Gerade in Statistikbüchern (vgl. z. B. Agresti und Finlay 2008) ist allerdings auch häufiger die abhängige Variable in den Spalten und die unabhängige in den Zeilen zu finden. Wichtig ist, dass die zur Beantwortung der Fragestellung richtige Prozentuierung berechnet wird.

In Tabelle 5.4 auf der gegenüberliegenden Seite ist eine Kreuztabellierung der Merkmale Schulabschluss (als Indikator für Bildung) und Wahlabsicht dargestellt. 367 Personen mit Hauptschulabschluss (HS) wollen die CDU/CSU wählen, 453 die SPD usw. Von den Befragten mit Realschulabschluss (RS) wollen 182 die CDU/CSU wählen und von den Befragten mit Abitur bzw. Fachhochschulreife (Abitur) äußern 119 eine Präferenz für die CDU. Am Ende jeder Zeile und jeder Spalte ist die Summe dieser Zeile bzw. Spalte wiedergegeben. Die Spalte, die mit „Summe" überschrieben ist, gibt die Häufigkeitsverteilung der abhängigen Variable an; diese wird auch als *Randverteilung* der abhängigen Variable bezeichnet. Die Randverteilung der unabhängigen Variable findet sich in der letzten, mit „Summe" beschriebenen Zeile. In der Ecke unten rechts steht die Gesamtsumme der Merkmalsträger, die in die Tabelle eingehen (2.218). Hier sind es die Befragten, die eine Partei genannt haben und für die der Schulabschluss bekannt ist.

Die absoluten Häufigkeiten lassen sich aber schlecht vergleichen. Absolut gesehen, wollen zwar erheblich mehr Befragte mit Hauptschulabschluss Christdemokraten und Christsoziale wählen (367) als Befragte mit Realschulabschluss (182) bzw. Abitur (119). Allerdings gibt es auch wesentlich mehr Befragte mit Hauptschulabschluss (1086) als Befragte mit Realschulabschluss (665) bzw. Befragte mit Abitur (467).

Übt die unabhängige Variable einen Einfluss auf die abhängige Variable aus, dann unterscheidet sich die *prozentuale Verteilung* der abhängigen Variablen für jede Ausprägung der unabhängigen Variablen.

Tabelle 5.4: Kreuztabelle der Wahlabsicht mit dem Schulabschluss – absolute Häufigkeiten

Wahlabsicht	Schulabschluss HS	RS	Abitur	Summe
CDU/CSU	367	182	119	668
SPD	453	244	131	828
FDP	77	71	49	197
Bündnis 90/Grüne	89	90	125	304
Republikaner	43	20	2	65
PDS	42	43	32	117
Andere Partei	15	15	9	39
Summe	1086	665	467	2218

Spalten-, Zeilen- und Totalprozente

Man muss also auch hier wieder die relativen Häufigkeiten bzw. Prozentwerte angeben. Dabei muss man beachten, dass sich in einer Kreuztabelle immer *drei verschiedene Prozentwerte* berechnen lassen: Zeilen-, Spalten- und Totalprozente. Zur Berechnung von **Zeilenprozenten** wird die Zeilensumme als Prozentuierungsbasis (= 100 %) genommen; bei **Spaltenprozenten** die Spaltensumme (= 100 %). Schließlich kann man die Häufigkeiten auf Basis der Gesamtsumme prozentuieren (**Totalprozente**). Damit sind drei inhaltlich völlig verschiedene Aussagen verbunden. Verwendet man die Zeilensumme als Prozentuierungsbasis, bezeichnet der Prozentwert einen Anteil an der Ausprägung des Merkmals in der Zeile. Verwendet man die Spaltensumme, bezeichnet der Prozentwert einen Anteil an der Ausprägung des Merkmals in der Spalte. Verwendet man die Gesamtsumme, bezeichnet der Prozentwert einen Anteil an allen Fällen, die in die Tabelle eingegangen sind.

In Tabelle 5.4 ließe sich der *Anteil der Befragten mit Hauptschulabschluss an allen potentiellen CDU/CSU-Wählern* mit $367/668 \times 100 = 54{,}9\,\%$ berechnen (Zeilenprozente). Der *Anteil der potentiellen CDU/CSU-Wähler an allen Befragten mit Hauptschulabschluss* berechnet sich dagegen mit $367/1086 \times 100 = 33{,}8\,\%$ (Spaltenprozente). Von den Befragten mit Hauptschulabschluss wollen also 33,8 % die CDU/CSU wählen. 54,9 % der Befragten mit einer CDU/CSU-Wahlabsicht haben einen Hauptschulab-

schluss. Totalprozente werden nicht so häufig benötigt; von allen Befragten haben genau 16,5 % (367/2218 × 100) einen Hauptschulabschluss *und* eine Wahlabsicht für die CDU/CSU. Totalprozente werden vor allem berechnet, um Aussagen über die Stabilität eines Merkmals zwischen zwei Zeitpunkten zu treffen (vgl. Kapitel 2.4).

Zusätzlich zu den absoluten Häufigkeiten sind in Tabelle 5.5 die Spaltenprozente und in Tabelle 5.6 auf der gegenüberliegenden Seite die Zeilenprozente angegeben. Die Prozentwerte sind *kursiv* hervorgehoben. In der Summen*spalte* findet sich die univariate Verteilung des Merkmals, das in den Zeilen steht (hier: Wahlabsicht), in der Summen*zeile* des Merkmals, das in den Spalten steht (hier: Bildung). Die Prozentwerte in der Summenspalte in Tabelle 5.5 entsprechen der prozentualen Verteilung der abhängigen Variable – hier der Wahlabsicht.[3]

Tabelle 5.5: Kreuztabelle der Wahlabsicht mit Bildung – absolute Häufigkeiten und Spaltenprozente

	Schulabschluss							
Wahlabsicht	HS		RS		Abitur		Summe	
CDU/CSU	367	*33,8*	182	*27,4*	119	*25,5*	668	*30,1*
SPD	453	*41,7*	244	*36,7*	131	*28,1*	828	*37,3*
FDP	77	*7,1*	71	*10,7*	49	*10,5*	197	*8,9*
Bündnis 90/Grüne	89	*8,2*	90	*13,5*	125	*26,8*	304	*13,7*
Republikaner	43	*4,0*	20	*3,0*	2	*0,4*	65	*2,9*
PDS	42	*3,9*	43	*6,5*	32	*6,9*	117	*5,3*
Andere Partei	15	*1,4*	15	*2,3*	9	*1,9*	39	*1,8*
Summe	1086	*100,0*	665	*100,0*	467	*100,0*	2218	*100,0*

Prozentuiert man spaltenweise, dann muss man zeilenweise interpretieren: Wie man anhand der Spaltenprozentwerte in der Tabelle 5.5 sieht, wollen 33,8 % der Befragten mit Hauptschulabschluss CDU/CSU wählen, aber nur 25,5 % der Befragten mit Abitur. Bündnis 90/Grüne wollen 26,8 % der Befragten mit Abitur wählen, aber nur 8,2 % der Befragten mit Hauptschulabschluss und 13,5 % der Befragten mit Realschulabschluss. Der Pro-

3 Die Prozentwerte unterscheiden sich von den in Tabelle 5.2 auf Seite 103 angegebenen, weil wir von den 2.298 Befragten, die eine Wahlabsicht angegeben haben, nur von 2.218 Personen den Schulabschluss kennen.

zentsatz der Befragten mit Abitur, die Bündnis 90/Grüne angeben, ist also verglichen mit dem Anteil, den die Grünen bei allen Befragten erzielen (13,7 %), überdurchschnittlich hoch. Gerade umgekehrte Verhältnisse zeigt die Wahlabsicht zugunsten der Republikaner. 4 % der Befragten mit Hauptschulabschluss wollen diese Partei wählen, aber nur 0,4 % der Befragten mit Abitur, wobei die Republikaner bei allen Befragten 2,9 % erzielen.

Bei zeilenweiser Prozentuierung wird spaltenweise interpretiert: Der Summenzeile von Tabelle 5.6 kann man entnehmen, dass insgesamt 49 % der Befragten einen Hauptschulabschluss haben, 30 % einen Realschulabschluss und 21,1 % Abitur. Hauptschüler sind unter den Wählern der CDU/CSU (54,9 %), der SPD (54,7 %) und der Republikaner (66,2 %) überproportional vertreten. Die Wählerschaft der Grünen weist dagegen mit 41,1 % einen stark überdurchschnittlichen Anteil an Befragten mit Abitur auf.

Tabelle 5.6: Kreuztabelle der Wahlabsicht mit dem Schulabschluss– absolute Häufigkeiten und Zeilenprozente

	Schulabschluss							
Wahlabsicht	HS		RS		Abitur		Summe	
CDU/CSU	367	*54,9*	182	*27,2*	119	*17,8*	668	*100,0*
SPD	453	*54,7*	244	*29,5*	131	*15,8*	828	*100,0*
FDP	77	*39,1*	71	*36,0*	49	*24,9*	197	*100,0*
Bündnis 90/Grüne	89	*29,3*	90	*29,6*	125	*41,1*	304	*100,0*
Republikaner	43	*66,2*	20	*30,8*	2	*3,1*	65	*100,0*
PDS	42	*35,9*	43	*36,8*	32	*27,4*	117	*100,0*
Andere Partei	15	*38,5*	15	*38,5*	9	*23,1*	39	*100,0*
Summe	1086	*49,0*	665	*30,0*	467	*21,1*	2218	*100,0*

Zur Präsentation von Tabellen gibt es keine einheitlichen Regeln. Das Layout der Tabelle sollte so gehalten sein, dass im Tabellenkopf die Spalten bezeichnet werden und am linken Tabellenrand die Zeilen. Je nach inhaltlicher Interpretation werden Zeilen- oder Spaltenprozente angegeben. Möchte man beide verwenden, so bietet es sich an, nur eine Tabelle zu erstellen, in der sowohl die absoluten Häufigkeiten als auch Zeilen- und Spaltenprozente stehen. Aus der Tabelle muss zudem die Größe und die Art der Prozentuierungsbasis hervorgehen.

5.2 Graphiken

5.2.1 Unterschiedliche Arten graphischer Darstellungen

Säulen- und Balkendiagramme

Bei einem Säulendiagramm werden die Daten durch vertikale (stehende) Rechtecke („Säulen“) wiedergegeben, beim Balkendiagramm durch horizontale (liegende) Rechtecke („Balken“).

Bei einem Balkendiagramm werden die Ausprägungen des Merkmals auf der vertikalen Achse abgetragen. Die Länge der mittig über den Kategorien eingezeichneten Balken entspricht den absoluten Häufigkeiten (relativen Häufigkeiten oder Prozentwerten) der Merkmalsausprägungen. Zwischen den Balken bleibt Platz. In Abbildung 5.1 wurde die schon bekannte Frage nach der Wahlabsicht als Balkendiagramm dargestellt. Die Länge der Balken entspricht den prozentualen Anteilen der einzelnen Parteien. Da es sich hier um ein nominal skaliertes Merkmal handelt, ist die Anordnung der Balken beliebig.

Abbildung 5.1: Balkendiagramm der Wahlabsicht

Quelle: ALLBUS 1994, n=2.298

Anstelle von Balkendiagrammen kann man ebenso gut Säulendiagramme verwenden. Ein Säulendiagramm der Wahlabsicht ist in Abbildung 5.2 auf der nächsten Seite wiedergegeben. Hier werden die Ausprägungen auf der horizontalen Achse abgetragen.

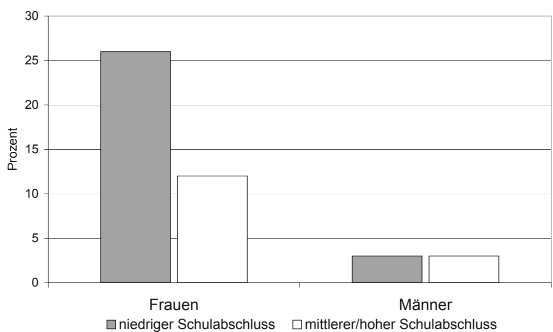
Abbildung 5.2: Säulendiagramm der Wahlabsicht

Quelle: ALLBUS 1994, n=2.298

Balken- und Säulendiagramme sind vor allem zur Darstellung nominal- und ordinal skalierter Variablen geeignet. Bei ordinal skalierten Merkmalen muss allerdings die Reihenfolge der Balken bzw. Säulen die Rangordnung der Merkmalsausprägungen wiedergeben. Die Breite der Balken oder Säulen ist beliebig, da diese nicht interpretiert werden kann. Balken- und Stabdiagramme eignen sich ebenfalls zur Darstellung diskreter, metrischer Merkmale mit einer überschaubaren Zahl von Ausprägungen, wie z. B. die Zahl der Kinder.

Mit Balkendiagrammen lässt sich auch der Zusammenhang zwischen zwei (und mehr) Merkmalen darstellen. In Abbildung 5.3 ist der Zusammenhang zwischen der Schulbildung und einer geringfügigen Beschäftigung getrennt für Männer und Frauen visualisiert. Man sieht deutlich, dass das Niveau der Schulbildung bei Frauen einen Einfluss auf eine geringfügige Erwerbstätigkeit hat, nicht jedoch bei Männern. Weil die entsprechenden Fallzahlen Tabelle 2.1 (S. 28) entnommen werden können, wurde hier auf eine Angabe verzichtet. Steht die Tabelle alleine, müsste hier eigentlich die Zahl der Frauen mit geringer Schulbildung, die Zahl der Frauen mit mittlerer/höherer Schulbildung, die Zahl der Männer mit geringer Schulbildung und die Zahl der Männer mit mittlerer/höherer Schulbildung angegeben werden, weil diese der Prozentuierung zugrunde liegen.

Abbildung 5.3: Mini-/Midi-Job nach Schulabschluss und Geschlecht

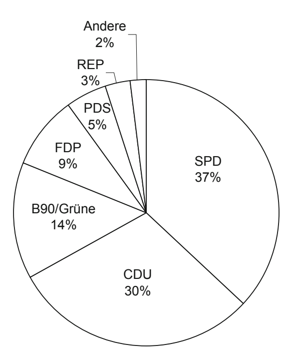


Quelle: SOEP, Welle W (gewichtet)

Kreisdiagramme

In Abbildung 5.4 sind die Angaben zur Wahlabsicht als Kreisdiagramm dargestellt. Die Größe der Kreissegmente ist proportional zur relativen Häufigkeit der jeweiligen Merkmalsausprägung. Kreisdiagramme eignen sich vor allem zur Illustration der Verteilung nominaler Merkmale.

Abbildung 5.4: Kreisdiagramm der Wahlabsicht (gültige Stimmen in Prozent)



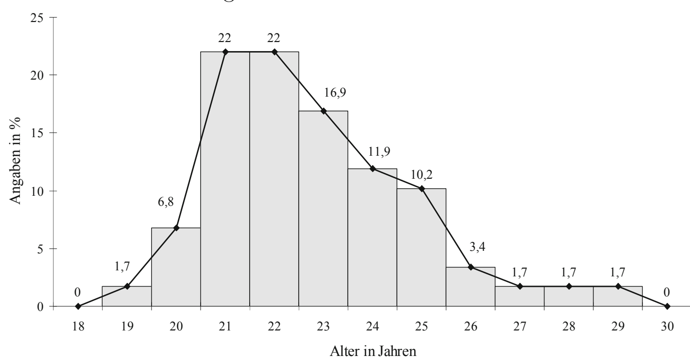
Quelle: ALLBUS 1994, n=2.298

Histogramme und Linienzüge

Histogramme und Linienzüge dienen der Darstellung stetiger Merkmale. Sie werden aber auch bei diskreten metrischen Merkmalen eingesetzt, die sehr viele Ausprägungen annehmen können, wie z. B. Einkommen.

Für die graphische Darstellung wird das Merkmal zunächst in benachbarte Klassen zusammengefasst. Bei *Histogrammen* werden die Messwerte durch Rechtecke über den Klassen symbolisiert, die unmittelbar aneinander angrenzen. Dies ist auch der auffälligste Unterschied zu Balken- und Säulendiagrammen. Die *Fläche* über den Klassen (Höhe x Breite der Rechtecke) ist proportional zu den absoluten bzw. relativen Häufigkeiten. Verbindet man die Mittelpunkte der Rechteckoberkanten durch Linien, dann erhält man einen Polygonzug. Für Linienzüge bzw. Polygone gilt das soeben Gesagte. Auch hier gibt die Fläche unter dem Linienzug Auskunft über die Häufigkeit/Anteile der Messwerte. Ob man sich für einen Linienzug oder ein Histogramm entscheidet, ist reine Geschmacksache. In Abbildung 5.5 wird das Alter der Teilnehmer zweier Statistik-Kurse sowohl durch ein Histogramm als auch durch einen Linienzug dargestellt.

Abbildung 5.5: Alter von Kursteilnehmern



Quelle: eigene Umfrage, n=59

Will man zwei verschiedene Verteilungen in einer Graphik darstellen, so bietet es sich an, ein Merkmal durch ein Histogramm, das andere durch einen Polygonzug darzustellen. Linienzüge eignen sich besonders für Zeitreihenanalysen. Ein Beispiel haben wir bereits im Kapitel 2.4 ken-

nengelernt, wo die Entwicklung der Anteile der Personen, die sich mit einer Partei identifizieren, für Ost- und Westdeutschland dargestellt wurde.

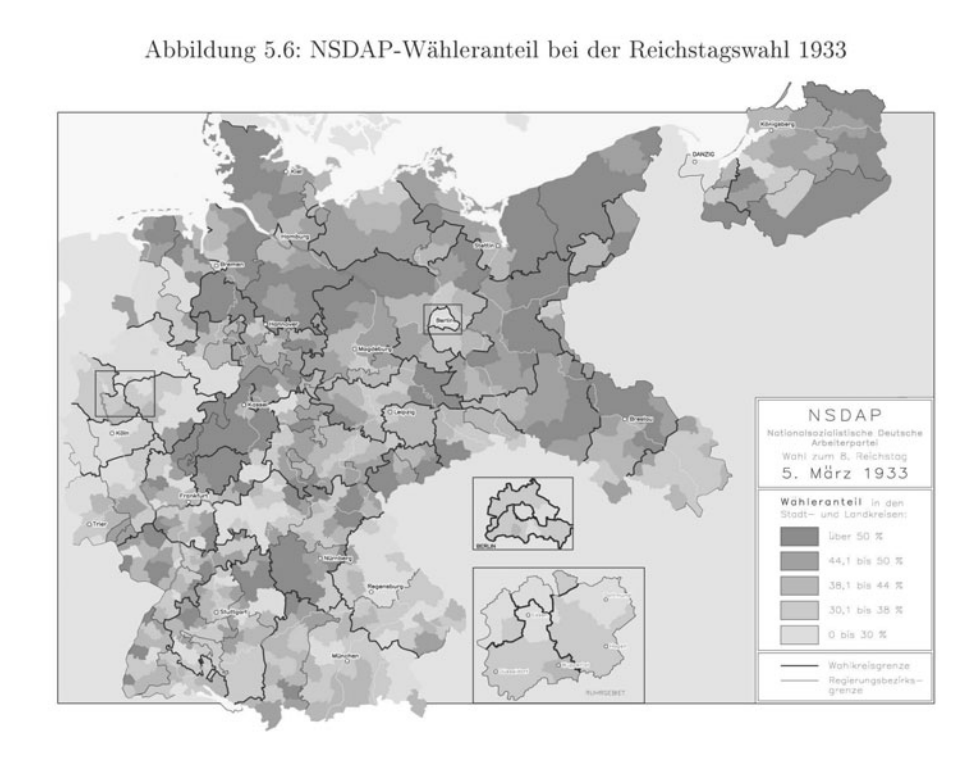
Zur Darstellung nominaler und ordinaler Merkmale sollten die Rechtecke nicht aneinander gezeichnet werden, weil dies suggerieren würde, dass Differenzen zwischen Merkmalsausprägungen interpretierbar seien. In Kapitel 3.3.1 wurde von dieser Regel abgewichen (Abbildung 3.3), weil bei den vier Items metrisches Messniveau unterstellt wurde. Wir werden in Kapitel 7.6 Streudiagramme kennen lernen, mit deren Hilfe der Zusammenhang zwischen zwei metrischen Merkmalen dargestellt werden kann.

Kartogramme

In Kartogrammen werden Merkmale geographischer Einheiten abgetragen. Die einzelnen Merkmalsausprägungen werden dabei durch unterschiedliche Schraffuren oder Farben repräsentiert.

Die Karte in Abbildung 5.6 auf der gegenüberliegenden Seite zeigt den Wähleranteil, den die NSDAP bei den Reichstagswahlen am 5. März 1933 in den einzelnen Stadt- und Landkreisen erzielen konnte. Je dunkler die Schraffur, umso höher der NSDAP-Anteil. Ganz dunkel sind die Stadt- und Landkreise dargestellt, in denen die NSDAP mehr als 50 % der Stimmen erhielt; ganz hell die Stadt- und Landkreise, in denen der NSDAP-Anteil unter 30 % lag. Zur Darstellung mussten die Prozentwerte gruppiert werden. Würde man jedem Prozentwert eine unterschiedliche Schraffur zuweisen, dann wäre das Kartogramm nicht mehr interpretierbar. Die Wahl der Kategoriegrenzen beeinflusst das Aussehen des Kartogramms natürlich entscheidend, was bei der Interpretation einer solchen Graphik berücksichtigt werden muss.

Abbildung 5.6: NSDAP-Wähleranteil bei der Reichstagswahl 1933



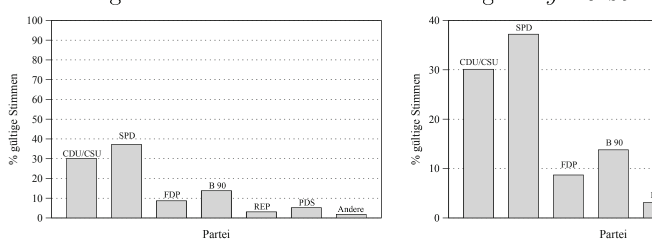
Diese Karte aus dem Projekt „Sozial- und Wahlatlas des Deutschen Reiches" wurde freundlicherweise von Dr. Jürgen Winkler zur Verfügung gestellt.

5.2.2 Missbrauch graphischer Darstellungen

Manchmal werden graphische Darstellungen bewusst oder unbewusst so gewählt, dass die eigene Interpretation der Daten gestützt wird. Meistens handelt es sich um Fahrlässigkeiten bei der Erstellung der Graphiken, manchmal jedoch auch um bewusste Manipulationen. Anhand des schon bekannten Beispiels zur Wahlabsicht wollen wir die Auswirkungen falscher Darstellungsweisen demonstrieren.

In Abbildung 5.7 ist dargestellt, wie sich eine Veränderung des Maßstabs der Einheiten der y-Achse auf die Aussagekraft der Graphik auswirkt. Eigentlich wäre die Skalierung nur dann korrekt, wenn sie von 0 % bis 100 % gehen würde, da ja *theoretisch* eine Partei 100 % der Stimmen bekommen kann. Außerdem wäre dann leicht erkennbar, wie viel eine Partei vom „ganzen Kuchen“ bekommen hat. Diese Darstellung ist im linken Diagramm in Abbildung 5.7 zu sehen. Der Nachteil dieser Darstellung besteht darin, dass die Unterschiede zwischen den Parteien nicht sehr deutlich ausfallen, da keine Partei mehr als 40 % Stimmen erhält und damit nicht einmal die Hälfte der Höhe der Graphik ausgeschöpft wird.

Abbildung 5.7: Wahlabsicht bei Veränderung des y-Achsen-Maßstabes

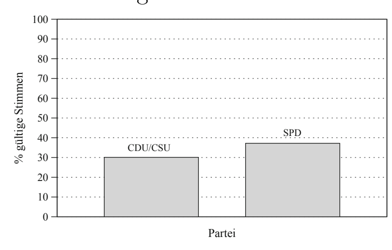


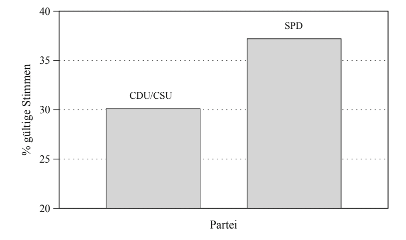
Quelle: ALLBUS 1994, n=2298

Um die Unterschiede hervorzuheben, könnte man deshalb den umgekehrten Weg einschlagen und die Skalierung nur von 0 % bis 40 % vornehmen, wie in der rechten Graphik in Abbildung 5.7 zu sehen ist. Die Unterschiede zwischen den Parteien werden dadurch stärker hervorgehoben. Die in Abbildung 5.2 auf Seite 111 gewählte Skalierung kann als ein Kompromiss zwischen diesen beiden extremen Darstellungen angesehen werden.

Falsch wäre es, wenn die Größenskala nicht bei null begänne. In Abbildung 5.8 wird dies anhand der Stimmanteile für CDU und SPD verdeutlicht. Während in der linken Graphik die korrekte Darstellung benutzt wurde, zeigt die rechte Graphik lediglich den Achsenausschnitt zwischen 20 % und 40 %. Dadurch werden die Größenverhältnisse zwischen CDU/CSU und SPD dramatisiert.

Abbildung 5.8: Wahlabsicht mit korrekter und falscher Grundlinie





Quelle: ALLBUS 1994, n=2.298

Zahlreiche Beispiele für Manipulationen graphischer Darstellungen finden sich bei Krämer (1991), wie man es richtig machen sollte bei Krämer (1994).

Aufgaben zu Tabellen und Graphiken

1. In der folgenden Tabelle ist das Wahlergebnis der Reichstagswahl vom 14. September 1930 wiedergegeben. Bitte ermitteln Sie die Wahlbeteiligung und den Anteil der ungültigen Stimmen. Prozentuieren Sie die Stimmen für die einzelnen Parteien (a) auf Basis der gültigen Stimmen und (b) auf Basis der Wahlberechtigten. Warum wird in der Regel auf gültige Stimmen prozentuiert?

Tabelle 5.7: Ergebnis der Reichstagswahl vom 14. September 1930

	Wahlergebnis
Wahlberechtigte	42.957.675
Abgegebene Stimmen	35.225.758
Ungültige Stimmen	254.901
Gültige Stimmen	34.970.857
KPD	4.592.090
USPD	11.902
SPD	8.577.738
DDP	1.322.385
Zentrum	4.127.910
BVP	1.059.141
DVP	1.659.774
DNVP	2.458.246
NSDAP	6.409.610
Sonstige	4.752.061

2. Bei der Reichstagswahl 1932 erzielte die NSDAP 37,3 % der gültigen Stimmen. Um wie viel Prozent und um wie viele Prozentpunkte stieg der Anteil der Nationalsozialisten im Vergleich zur Reichstagswahl von 1930?
3. Bitte stellen Sie das Wahlergebnis der Reichstagswahl 1930 graphisch dar! Welche Diagramme können zur Darstellung verwandt werden?

4. Auf die im ALLBUS 1994 gestellte Frage, *„Und wie wird Ihre eigene wirtschaftliche Lage in einem Jahr sein?“*, konnten die Befragten von *„wesentlich besser“* bis *„wesentlich schlechter“* antworten. In Tabelle 5.8 sind die Antworten getrennt für west- und ostdeutsche Befragte wiedergegeben:

   Tabelle 5.8: Wirtschaftliche Einstellungen im ALLBUS 1994

	West	Ost	Summe
wesentlich besser	38	22	60
etwas besser	348	254	602
gleichbleibend	1588	661	2249
etwas schlechter	293	119	412
wesentlich schlechter	23	21	44
Summe	2290	1077	3367

   Bitte berechnen Sie Spalten-, Zeilen- und Totalprozente. Interpretieren Sie die inhaltliche Aussage der Tabelle!

6 Lage- und Streuungsmaße

Im vorangegangenen Kapitel wurden Häufigkeitsverteilungen sowie deren Darstellung durch Tabellen und Graphiken behandelt. In diesem Kapitel werden nun *statistische Maßzahlen* vorgestellt, die die zentrale Lage einer Verteilung und die Streuung der Messwerte charakterisieren.

In Tabelle 6.1 ist die Studiendauer von 11 Absolventen der Politikwissenschaft wiedergegeben (das Beispiel ist fiktiv). In der linken Tabelle liegen die Messwerte der einzelnen Personen als *Urliste* vor, d. h. so, wie wir sie willkürlich nacheinander notiert haben. Um die Übersichtlichkeit zu erhöhen, wurden die Messwerte nach ihrer Größe geordnet. Diese so genannte *primäre Tafel* ist in der rechten Tabelle wiedergegeben.

Tabelle 6.1: Semesterzahl von Politologen: ungruppierte Daten

Urliste

i	x_i
1	12
2	14
3	10
4	15
5	11
6	20
7	12
8	12
9	11
10	13
11	13

primäre Tafel

i	x_i
1	10
2	11
3	11
4	12
5	12
6	12
7	13
8	13
9	14
10	15
11	20

Mit i wird der Laufindex für die einzelnen Merkmalsträger (hier also Personen) bezeichnet, mit x_i die konkrete Merkmalsausprägung des i-ten Merkmalsträgers (bzw. der i-ten Person). Allgemein nimmt der Laufindex i

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen*,
DOI 10.1007/978-3-531-91879-2_6,

Werte zwischen 1 und n an, wobei n der Anzahl der Merkmalsträger (Personen) entspricht. Im Beispiel „läuft“ der Index i also von 1 bis 11, da die Messwerte – die Semesterzahl – bei 11 Personen erhoben wurden. Mit $i = 4$ wird also der vierte Messwert bezeichnet, mit x_4 die konkrete Merkmalsausprägung der vierten Person. In der Urliste nimmt x_4 den Wert 15 an, d. h. diese Person hat bis zur Magisterprüfung 15 Semester lang studiert. In der primären Tafel nimmt dagegen x_4 den Wert 12 an, d. h. der Laufindex wird bei der Sortierung nicht berücksichtigt.

Sowohl bei der Urliste als auch bei der primären Tafel werden die Messwerte einzeln aufgeführt, es handelt sich daher um *ungruppierte Daten*. In einer Häufigkeitstabelle (Kapitel 5.1.1) werden dagegen gleiche Messwerte zusammengefasst (*gruppierte Daten*). Zwischen ungruppierten und gruppierten Daten gibt es keinen Informationsverlust, da die Merkmalsausprägung einer jeden Person vollständig reproduzierbar ist.

Tabelle 6.2: Semesterzahl von Politologen: Häufigkeitstabelle

k	x_k	f_{x_k}	%	kum. %
1	10	1	9,1 %	9,1 %
2	11	2	18,2 %	27,3 %
3	12	3	27,3 %	54,6 %
4	13	2	18,2 %	72,8 %
5	14	1	9,1 %	81,9 %
6	15	1	9,1 %	91,0 %
7	20	1	9,1 %	100,1 %
$\sum$		11	100,1 %	100,1 %

Der Laufindex für die einzelnen Kategorien wird mit k bezeichnet und „läuft“ von 1 bis m, wobei m der Zahl der Kategorien entspricht; in diesem Beispiel sind es sieben. *Der Laufindex für Kategorien k sollte auf keinen Fall mit dem Laufindex für Merkmalsträger (hier: Personen) i verwechselt werden.* Die Merkmalsausprägung einer Kategorie wird mit x_k bezeichnet, x_4 entspricht der Merkmalsausprägung 13 Semester. Da die Daten zusammengefasst wurden, benötigen wir noch eine Angabe über die Häufigkeit f_{x_k} mit der die Merkmalsausprägungen auftreten. f_{x_4} ist hier 2, was bedeutet, dass zwei Politologen 13 Semester bis zum Abschluss benötigten. Prozentual ausgedrückt haben 18,2 % der Studierenden (2 von

11) 13 Semester bis zum Abschluss des Studiums benötigt. In der letzten Spalte sind die kumulierten (addierten) Prozente angegeben. 9,1 % der Studierenden haben 10 Semester bis zum Abschluss benötigt, 27,3 % der Studierenden haben 11 oder 10 Semester benötigt, 54,6 % der Studierenden nicht mehr als 12 Semester, 73 % weniger als 13 Semester usw. Die Summe der Prozente addiert sich wegen Rundungsfehlern hier nicht ganz exakt zu 100 %. Die Berechnung kumulierter Prozentwerte ist erst ab ordinalskaliertem Skalenniveau sinnvoll, weil die Merkmalsausprägungen dazu nach der Größe sortiert werden müssen.

6.1 Lagemaße

Mittelwerte kennzeichnen die zentrale Lage einer Verteilung. Wenn vom Mittelwert gesprochen wird, dann ist in der Regel ein spezieller Mittelwert, nämlich das arithmetische Mittel, gemeint. Die drei wichtigsten Mittelwerte sind:

1. Modalwert
2. Median
3. Arithmetisches Mittel

Welchen der drei Mittelwerte man berechnet, hängt zum einem vom *Skalenniveau des Merkmals* und zum anderen von der zu treffenden inhaltlichen Aussage ab. Bei nominal skalierten Merkmalen kommt der Modalwert in Frage, bei ordinalen Merkmalen zusätzlich der Median und bei metrischen Merkmalen lässt sich auch das arithmetische Mittel sinnvoll interpretieren.

Zwei weitere Mittelwerte für mindestens ratioskalierte Merkmale, das *geometrische* und das *harmonische Mittel*, sind für uns von untergeordneter Bedeutung (vgl. Sachs 2006, S. 76–78). Das geometrische Mittel kommt bei positiven, ratioskalierten Daten zum Einsatz und ist inhaltlich bei der Berechnung durchschnittlicher Wachstumsfaktoren (Umsatz, Zinsen usw.) angemessen.

6.1.1 Modalwert

Der **Modalwert** ist der **Messwert**, der in einer Verteilung am **häufigsten** vorkommt. Bei einer graphischen Darstellung ist der Modalwert also

der Gipfel bzw. das Maximum der Verteilung. Die Bezeichnung für den Modalwert ist nicht einheitlich. Wir benutzen x_{Mo}.

Kommen zwei Messwerte in einer Verteilung (annähernd) gleich häufig vor, dann kann man zwei Modalwerte angeben. Sind die beiden häufigsten Messwerte nicht benachbart, dann spricht man von einer *bimodalen Verteilung*. Sind die beiden häufigsten Werte benachbart, dann kann man bei metrischen Merkmalen das arithmetische Mittel der beiden Modalwerte ausrechnen. Bei mehr als zwei Modalwerten wird im Allgemeinen auf deren Angabe verzichtet.

Modalwerte haben den Vorteil, dass sie direkt aus der Verteilung ersichtlich sind. Der Modalwert der Studiendauer (vgl. Tabelle 6.2) beträgt 12, da dies der Wert der am stärksten besetzten Kategorie ($f_{x_k} = 3$) ist. (Das heißt aber nicht, dass *die meisten* der elf Politologen 12 Semester bis zum Studienabschluss benötigt haben.) Modalwerte lassen sich für alle Messniveaus bestimmen. Der Modalwert der Religionszugehörigkeit (nominales Merkmal, Tabelle 6.3) ist bei westdeutschen Befragten „Evangelisch/Freikirche", bei ostdeutschen Befragten „Keine Konfession".

Tabelle 6.3: Religionszugehörigkeit

	Westdeutschland		Ostdeutschland	
	Häufigkeit	rel. Häufig.	Häufigkeit	rel. Häufig.
Evangelisch/Freikirche	905	0,40	284	0,25
Katholisch	838	0,37	45	0,04
Andere christl. Religion	56	0,02	15	0,01
Nicht christl. Religion	109	0,05	10	0,01
Keine Konfession	377	0,16	765	0,68
Gesamt	2285	1,00	1119	1,00

ALLBUS 2006

6.1.2 Median

Der **Median** ist der Wert, der die nach der Größe aufsteigend sortierten Messwerte in zwei Hälften teilt. Der Median ist also der Wert, der in der Mitte liegt. Der Median der drei Einkommen 1000 Euro, 1500 Euro und 8000 Euro ist 1500 Euro. Die Bezeichnung des Medians ist $\tilde{x}$.

Um den Median zu ermitteln, muss man die Messwerte zunächst ordnen, d. h. die primäre Tafel erstellen. Anschließend sucht man den Wert, der

in der Mitte liegt. Bei einer **ungeraden Zahl** von Messwerten existiert genau ein Messwert, der in der Mitte liegt, und zwar an der Stelle $\frac{n+1}{2}$. Der Median ist die Merkmalsausprägung des Messwerts, der an der $\frac{n+1}{2}$ten Stelle in der geordneten Verteilung liegt:

$$\tilde{x} = x_{\frac{n+1}{2}} \,. \tag{6.1}$$

Im Beispiel zur Studiendauer, die für $n = 11$ Studierende erhoben wurde, beträgt der Median

$$\tilde{x} = x_{\frac{11+1}{2}} = x_6 = 12\,\text{Semester.}$$

Bei 11 Messwerten liegt der sechste Messwert – $(11 + 1)/2 = 6$ – in der Mitte der Verteilung. Die Merkmalsausprägung des sechsten Wertes ist $x_6 = 12$ Semester. Die mittlere Studiendauer beträgt also 12 Semester. Die Hälfte der Studierenden benötigt bis zum Studienabschluss weniger oder gleich 12 Semester, die Hälfte mehr als 12 Semester. In der ersten Hälfte befinden sich bereits zwei Studierende, die ebenfalls 12 Semester studiert haben. Bei diesen beiden kann man nicht davon sprechen, dass sie „schneller" studiert haben.

Bei einer **geraden Zahl von Messwerten** existieren zwei mittlere Werte, und zwar an den Stellen $\frac{n}{2}$ und $\frac{n}{2} + 1$. Bei 12 Messwerten liegen der sechste ($\frac{12}{2}$) und der siebte ($\frac{12}{2} + 1$) Messwert in der Mitte. Es hat sich bei einer ungeraden Zahl von Messwerten eingebürgert, den Median als arithmetisches Mittel der Merkmalsausprägungen der beiden in der Mitte liegenden Messwerte zu berechnen:

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \,. \tag{6.2}$$

Würde in unserem Beispiel noch ein Politologe mit $x_{12} = 21$ Semestern hinzukommen (insgesamt sind es dann $n = 12$ Personen), würde sich der Median wie folgt ermitteln:

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_6 + x_7}{2} = \frac{12 + 13}{2} = \frac{25}{2} = 12{,}5\,.$$

Die Merkmalsausprägung des sechsten Messwerts (x_6) ist 12 Semester, die Merkmalsausprägung des siebten Messwerts (x_7) beträgt 13 Semester. Das arithmetische Mittel aus diesen beiden Werten ist 12,5 Semester. Die mittlere Studiendauer beträgt nun also 12,5 Semester. Alternativ können auch die beiden mittleren Werte angegeben werden, was bei ordinalskalierten Merkmalen angemessener ist.

Die Berechnung des Medians setzt lediglich voraus, dass die Messwerte in eine Reihenfolge gebracht werden können. Er ist deshalb für alle Daten angemessen, die mindestens ordinalskaliert sind. Bei einer großen Zahl von Beobachtungen lässt sich der Median am einfachsten aus der Häufigkeitstabelle ermitteln. Tabelle 6.4 enthält die Verteilung der schulischen Abschlüsse der westdeutschen Befragten des ALLBUS 2006. Insgesamt liegen für 2.221 Personen Beobachtungen vor. Der Median ist daher der $(n+1)/2 = (2221+1)/2 = 1.111$te Messwert, wenn die Schulabschlüsse wie in der Tabelle nach der Höhe des Abschlusses sortiert sind. In die Kategorie Hauptschule (niedrigster Abschluss) fallen 961 Beobachtungen. Die Kategorien Hauptschulabschluss und Mittlere Reife umfassen zusammen $(961 + 660) = 1621$ Beobachtungen usw. Die 962te bis 1621te Beobachtung fällt in die Kategorie Mittlere Reife, die damit auch die 1.111te Beobachtung beinhaltet. Der Median ist daher $x_{1111} =$ ‚Mittlere Reife'. Am leichtesten lässt sich der Median aus der Spalte der kumulierten Prozentwerte ablesen. Der Median ist der Wert, an dem 50 % der Beobachtungen einen kleineren Wert haben. 50 % (letzte Spalte) werden in der Kategorie Mittlere Reife erreicht.

Tabelle 6.4: Schulabschluss

x_k	f_{x_k}	%	kum. %
Hauptschule	961	43 %	43 %
Mittlere Reife	660	30 %	73 %
Fachhochschulreife	148	7 %	80 %
Hochschulreife	452	20 %	100 %
Gesamt	2221	100 %	100 %

ALLBUS 2006, westdeutsche Befragte

6.1.3 Arithmetisches Mittel

Das **arithmetische Mittel** ist der Wert, den alle Merkmalsträger, also z. B. Personen, **im Durchschnitt** aufweisen. Landläufig bezeichnet man das arithmetische Mittel deshalb auch als Durchschnittswert. Das arithmetische Mittel wird mit $\bar{x}$ bezeichnet. Die Berechnung des arithmetischen Mittels setzt mindestens intervallskalierte Daten voraus.

- Bei *ungruppierten Daten* – wie in Tabelle 6.1 – berechnet sich das arithmetische Mittel wie folgt:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \,. \tag{6.3}$$

  Zur Berechnung des arithmetischen Mittels werden die Merkmalsausprägungen aller Merkmalsträger summiert ($\sum x_i$) und anschließend durch die Anzahl der Merkmalsträger (n) dividiert (vgl. zum Rechnen mit dem Summenzeichen Bortz 2004, S. 703 f.). Eine Division durch n bedeutet immer, dass ein Durchschnittswert ausgerechnet wird. Hier ist es also die durchschnittliche Ausprägung des Merkmals x_i, im konkreten Beispiel also die durchschnittliche Semesterzahl:

$$\begin{aligned}
\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\
&= \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11}}{n} \\
&= \frac{10 + 11 + 11 + 12 + 12 + 12 + 13 + 13 + 14 + 15 + 20}{11} \\
&= \frac{143}{11} = 13\,\text{Semester.}
\end{aligned}$$

  Die durchschnittliche Studiendauer beträgt also 13 Semester. Ob man zur Berechnung der Summe der Merkmalsausprägungen die Werte der primären Tafel oder der Urliste entnimmt, ist natürlich völlig egal.
- Für Daten, die in einer Häufigkeitstabelle vorliegen (*gruppierte Daten*) – wie in Tabelle 6.2 –, sieht die Formel etwas anders aus:

$$\bar{x} = \frac{\sum_{k=1}^{m}(x_k \cdot f_{x_k})}{n}. \tag{6.4}$$

Jetzt werden also nicht die Merkmalsausprägungen aller Personen (x_i) summiert, sondern die Merkmalsausprägungen der Kategorien (x_k) multipliziert mit deren Häufigkeit (f_{x_k}). Die Division durch n bleibt. Angewendet auf das Beispiel in Tabelle 6.2 berechnet sich $\bar{x}$:

$$\begin{aligned}\bar{x} &= \frac{\sum_{k=1}^{m}(x_k \cdot f_{x_k})}{n}\\ &= \frac{10 \cdot 1 + 11 \cdot 2 + 12 \cdot 3 + 13 \cdot 2 + 14 \cdot 1 + 15 \cdot 1 + 20 \cdot 1}{11}\\ &= \frac{143}{11} = 13\,\text{Semester.}\end{aligned}$$

Das Ergebnis ist natürlich dasselbe.

Das arithmetische Mittel ist der am häufigsten verwendete Mittelwert. Er nutzt alle beobachteten Informationen aus. Im Gegensatz zum Median wird das arithmetische Mittel allerdings durch einen oder mehrere stark von den restlichen Werten abweichende Werte – so genannte „Ausreißer“ – verzerrt. Dies kann man sich an unserem Beispiel verdeutlichen, in dem es einen Wert gibt, der deutlich von den anderen Werten abweicht ($x_{11} = 20$ Semester). Berechnet man nun Modalwert, Median und arithmetisches Mittel für die Gesamtverteilung sowie für eine Verteilung, in der dieser „Ausreißer“ weggelassen wird, kommt man zu folgendem Ergebnis:

Tabelle 6.5: Einfluss von Ausreißern

	alle Messwerte	Messwerte ohne x_{11}
x_{Mo}	12	12
$\tilde{x}$	12	12
$\bar{x}$	13	12,3

Während Modalwert und Median unverändert bleiben, verkleinert sich das arithmetische Mittel in der Verteilung ohne den Ausreißer. Die durchschnittliche Studiendauer beträgt nun nicht mehr 13 Semester, sondern 12,3 Semester. *Das arithmetische Mittel wird durch den extremen Wert verzerrt*, während Modalwert und Median gleich bleiben.

Das arithmetische Mittel weist zwei Eigenschaften auf, die man sich bei anderen statistischen Berechnungen – zum Beispiel der im Anschluss behandelten Streuungsmaße – zunutze machen kann:

1. Die Summe der Abweichungen aller Messwerte vom Mittelwert ist 0. Mathematisch ausgedrückt, sieht das folgendermaßen aus:

   $$\sum_{i=1}^{n}(x_i - \bar{x}) = 0\,.$$

   Würde man in diese Formel irgendeinen anderen Wert anstelle des arithmetischen Mittels einsetzen, würde ein Wert $\neq 0$ herauskommen, d. h. dass nur das arithmetische Mittel diese Eigenschaft besitzt.
2. Die *Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert* bzw. die „Summe der Abweichungsquadrate“ (SAQ) ist minimal. Auch dazu wieder der mathematische Ausdruck:

   $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \text{min.}$$

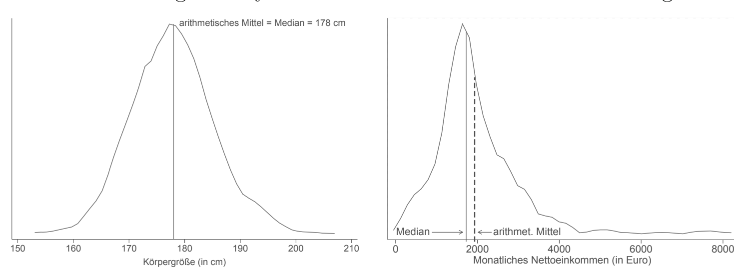
   „Minimal“ in dieser Formel heißt, dass bei der Berechnung der quadrierten Abweichungen der Messwerte von irgendeinem anderen Wert das Ergebnis auf jeden Fall größer wäre als bei Verwendung des arithmetischen Mittels.

Wir wissen nun, dass die durchschnittliche Studiendauer $\bar{x}$ der elf Politologen 13 Semester beträgt, die mittlere Studiendauer ($\tilde{x}$) bei 12 Semestern liegt, und dass die am stärksten besetzte Kategorie (x_{Mo}) 12 Semester ist. Das arithmetische Mittel ist der größte Wert, gefolgt von Median und Modalwert.

In symmetrischen Verteilungen sind Median und arithmetisches Mittel identisch, wie man an der Verteilung der Körpergröße der im ALLBUS

2004 in Westdeutschland befragten Männer (linker Teil in Abbildung 6.1) erkennen kann ($\tilde{x} = \bar{x} = 178$ cm). In schiefen Verteilungen wird der arithmetische Mittelwert stärker in Richtung des längeren Endes der Verteilung beeinflusst. Das klassische Beispiel für eine schiefe Verteilung sind Einkommen. Die Verteilung der monatlichen Nettoeinkommen der in Westdeutschland befragten Männer (ALLBUS 2006, rechter Teil in Abbildung 6.1) ist *linkssteil* (rechtsschief): Die Einkommensverteilung steigt zunächst (am linken Ende) steil an und fällt dann nach rechts flach ab. Das längere Ende der Verteilung ist bei den höheren Einkommen. Das arithmetische Mittel der Einkommen (gestrichelte Linie) wird durch die sehr hohen Einkommen nach oben beeinflusst, der Median (durchgezogene Linie) nicht. Die mittleren Einkommen ($\tilde{x}$) sind daher niedriger als die durchschnittlichen Einkommen ($\bar{x}$). Möchte man Einkommen hoch darstellen, so wird man mit den durchschnittlichen Einkommen $\bar{x}$ argumentieren, möchte man diese niedrig darstellen, wird man die mittleren Einkommen $\tilde{x}$ verwenden. Auch die Studiendauer der Politologen ist linkssteil verteilt: Der Median beträgt 12 Semester, das arithmetische Mittel 13 Semester (Tabelle 6.5). In einer *rechtssteilen* (linksschiefen) Verteilung befindet sich das längere Ende der Verteilung am linken Ende, was empirisch jedoch seltener vorkommt (vgl. Abbildung 3.3, S. 50). In rechtssteilen Verteilungen ist das arithmetische Mittel kleiner als der Median.

Abbildung 6.1: Symmetrische und linkssteile Verteilung



In unimodalen symmetrischen Verteilungen fallen alle drei Mittelwerte in einem Punkt zusammen ($x_{Mo} = \tilde{x} = \bar{x}$). In einer linkssteilen Verteilung ist der Modalwert der kleinste Wert, gefolgt von Median und arithmetischem Mittel ($x_{Mo} < \tilde{x} < \bar{x}$). In einer rechtssteilen Verteilung nimmt das arithmetische Mittel den kleinsten Wert, der Median den mittleren Wert und der Modalwert den größten Wert an ($\bar{x} < \tilde{x} < x_{Mo}$).

6.2 Streuungsmaße

Ein Mittelwert beschreibt die Verteilung umso besser, je näher die Daten beieinander liegen. Streuen die Daten jedoch stark, ist die Berücksichtigung eines *Streuungsmaßes* ratsam. Die Mittelwerte einer beliebigen Verteilung können nämlich bei unterschiedlicher Streuung identisch ausfallen. So ist z. B. in einer Verteilung von drei Einkommen in Höhe von 2000 Euro, 3000 Euro und 4000 Euro das leicht zu errechnende Durchschnittseinkommen 3000 Euro. Auch bei drei Einkommen in Höhe von 100 Euro, 900 Euro und 8000 Euro ergäbe sich ein arithmetisches Mittel von 3000 Euro. Die beiden Verteilungen streuen jedoch unterschiedlich stark.

Dargestellt werden die folgenden Maßzahlen:

1. Index qualitativer Variation
2. Variationsweite
3. Quartilabstand
4. Varianz
5. Standardabweichung
6. Variationskoeffizient

Welches Streuungsmaß angemessen ist, hängt auch hier vom Skalenniveau der betrachteten Merkmale ab.

6.2.1 Index qualitativer Variation

Nominale Streuungsmaße sind nicht sehr weit verbreitet. Die Maße, die zur Verfügung stehen – wie die Devianz (Kühnel und Krebs 2007, 96 ff.) oder der hier dargestellte Index qualitativer Variation (IQV) –, beruhen darauf, dass die Streuung bei nominalen Merkmalen maximal ist, wenn die Ausprägungen eines Merkmals gleich häufig besetzt sind. Die Streuung ist minimal, wenn alle Beobachtungen in eine Kategorie fallen. Bei einem Merkmal mit zwei Ausprägungen ist die Streuung maximal, wenn jede der beiden Kategorien 50 % der Beobachtungen beinhaltet. Sie ist minimal, wenn nur eine der beiden Kategorien besetzt ist.

Der Index qualitativer Variation berechnet sich nach

$$IQV = \frac{1 - \sum_{k=1}^{m} p_k^2}{(m-1)/m}\,, \tag{6.5}$$

wobei m die Zahl der Kategorien und p_k die relative Häufigkeit der k-ten Kategorie angibt.

Für die westdeutschen Befragten beträgt die Streuung der Religionszugehörigkeit (Tabelle 6.3, S. 123)

$$IQV = \frac{1 - (.4^2 + .37^2 + .02^2 + .05^2 + .16^2)}{4/5} = 0,84\,. \tag{6.6}$$

Für die ostdeutschen Befragten ist die Streuung der Religionszugehörigkeit geringer. Rund 70 % der Befragten befinden sich hier in einer einzigen Kategorie, nämlich „Keine Konfession". Der Index qualitativer Variation beträgt 0,37.

Bei einer gleichen Verteilung der Beobachtungen auf alle Kategorien $p_i = 1/K$ wird der Index 1 (maximale Streuung). Sofern eine Kategorie alle Beobachtungen beinhaltet ($p_k = 1$) nimmt der Index einen Wert von Null an (keine Streuung).

6.2.2 Variationsweite

Die **Variationsweite** (auch: Spannweite) gibt den Abstand zwischen dem maximalen und minimalen Wert einer Verteilung an. Die Bezeichnung ist normalerweise V, manchmal auch r, wegen der englischen Bezeichnung *range*. Letzteres wird von uns aber nicht empfohlen, da die Bezeichnung r in der Regel für den *Pearson'schen Korrelationskoeffizienten* verwendet wird (vgl. Kapitel 7.6). Die Berechnung von V setzt voraus, dass die Daten eine Rangordnung haben.

$$V = x_{max} - x_{min} \tag{6.7}$$

V ist der Wert mit der *größten* Merkmalsausprägung (nicht zu verwechseln mit dem Wert mit der *häufigsten* Merkmalsausprägung) abzüglich des Wertes mit der *kleinsten* Merkmalsausprägung (ebenso nicht zu verwechseln mit dem Wert mit der seltensten Merkmalsausprägung). Im Beispiel der Studiendauer ergibt sich

$$V = x_{max} - x_{min} = 20 - 10 = 10\,\text{Semester}.$$

Die Variationsweite ist 10. Zwischen dem Studierenden, der als Erster sein Studium beendete, und demjenigen, der zuletzt das Studium abschloss, liegen also 10 Semester.

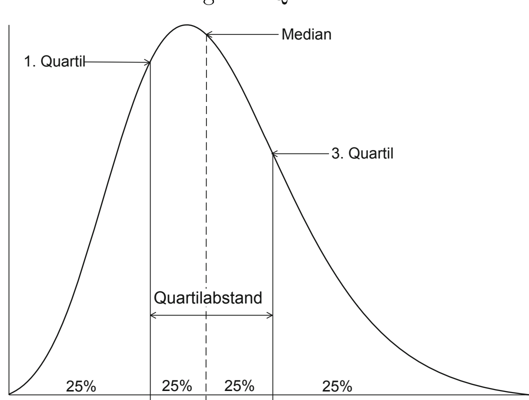
6.2.3 Quartilabstand

Die Variationsweite nutzt lediglich die beiden Werte an den Enden der Verteilung und ist daher empfindlich gegenüber Ausreißern. Die Spannweite der Semesterzahl beträgt 10; ohne den Studierenden, der 20 Semester studiert hat, betrüge die Variationsweite $(15 - 10) = 5$ Semester.

Der Quartilabstand ist nicht abhängig von den Werten an den Enden der Verteilung. Er gibt die Differenz zwischen dem 3. Quartil und dem 1. Quartil einer Verteilung an (vgl. Abbildung 6.2). Zur Bestimmung der Quartile werden die Messwerte aufsteigend sortiert und in vier gleich stark besetzte Gruppen geteilt. Jedes Quartil enthält 25 % der Messwerte. 25 % der Messwerte sind kleiner als oder gleich dem Wert des 1. Quartils, 75 % sind gleich groß oder größer. Das 2. Quartil ist der Median. Das 3. Quartil ist der Wert, an dem 75 % der Werte kleiner oder gleich groß sind und 25 % gleich groß oder größer. Die mittleren 50 % der Messwerte befinden sich zwischen dem 1. und 3. Quartil einer Verteilung.

Da die Zahl der Messwerte nicht immer exakt durch 4 teilbar ist, gibt es verschiedene Berechnungsmethoden. Die Werte des 1. und 3. Quartils können ebenso wie der Median (2. Quartil) einfach an den kumulierten Prozentwerten abgelesen werden (Tabelle 6.2, letzte Spalte). 25 % der Studierenden haben nicht länger als 11 Semester, 75 % nicht länger als 14 Semester studiert. Das 1. Quartil der Studienlänge der 11 Politologen liegt bei 11 Semestern und das 3. Quartil bei 14 Semestern. 50 % der Studierenden haben zwischen 11 und 14 Semestern bis zum Abschluss benötigt, der Quartilabstand beträgt $(14 - 11) = 3$ Semester.

Abbildung 6.2: Quartilabstand



Minimum, 1. Quartil, Median, 3. Quartil und Maximum werden häufig zur Charakterisierung der Lage und Breite einer Verteilung herangezogen und als *5-Punkte-Zusammenfassung einer Verteilung* (Tuckey 1977) bezeichnet.

Tabelle 6.6: Semesterzahl - 5 Punkte-Zusammenfassung

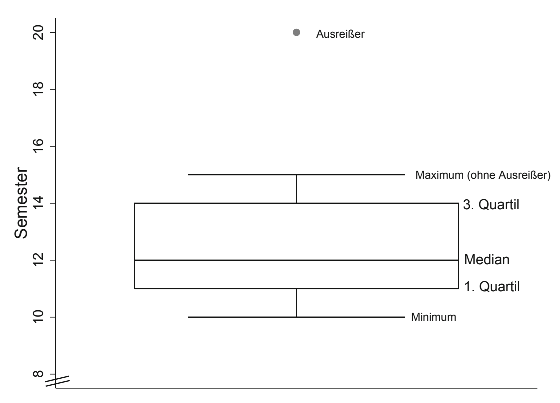
Minimum	10
1. Quartil	11
Median	12
3. Quartil	14
Maximum	20

Grafisch wird die 5-Punkte-Zusammenfassung einer Verteilung durch Box-and-Whisker-Plots visualisiert (Abbildung 6.3). Die untere Grenze der Box ist das 1. Quartil (11 Semester), die obere Grenze der Box ist das 3. Quartil (14 Semester). Die Länge der Box entspricht dem Quartilabstand $14 - 11 = 3$ Semester). Innerhalb der Box ist der Median (12 Semester) durch eine Linie gekennzeichnet. An der Box erkennt man deutlich, dass die Studiendauer linkssteil verteilt ist. Der Abstand zwischen Median und 3. Quartil ist größer als der Abstand zwischen Median und 1. Quartil.

Die Box wird durch zwei Zäune (*whisker*) nach oben und nach unten verlängert. Die Zäune entsprechen dem Minimum und Maximum der Vertei-

lung, sofern Minimum bzw. Maximum keine Ausreißer sind. Ausreißer sind Messwerte, die weiter als den 1,5fachen Quartilabstand von der Box entfernt sind. Sie werden einzeln dargestellt. Sind Ausreißer vorhanden, dann ist der Zaun an der Stelle des kleinsten bzw. größten Messwertes, der kein Ausreißer ist. Der Quartilabstand der Studiendauer beträgt 3 Semester, der 1,5fache Quartilabstand $1,5(3) = 4,5$ Semester. Die untere Grenze für Ausreißer ist demnach 1. Quartil $- 1,5(\text{IQR}) = 11 - 1,5(3) = 6,5$ Semester, die obere Grenze für Ausreißer beträgt 3. Quartil$+1,5(\text{IQR}) = 11 + 1,5(3) = 15,5$ Semester. Semesterzahlen kleiner als 6,5 und größer als 15,5 sind Ausreißer. Ausreißer nach unten, das heißt Studierende, die weniger als 6,5 Semester bis zum Abschluss benötigt haben, existieren nicht. Der untere Zaun entspricht deshalb dem Minimum der Verteilung (10 Semester). Nach oben existiert ein Ausreißer (Wert $>$ 15,5 Semester), nämlich 20 Semester. Er ist einzeln in der Abbildung visualisiert. Der obere Zaun ist der größte Wert der Verteilung, der kein Ausreißer ist, nämlich 15 Semester (vgl. Tabelle 6.1).

Abbildung 6.3: Box-and-Whisker-Plot



Bei intervall- und ratioskalierten Merkmalen gibt der Quartilabstand an, wie weit die mittleren 50 % der Messwerte einer Verteilung voneinander entfernt sind. Bei ordinalskalierten Merkmalen kann der Abstand zwischen den Quartilen nur im Sinne von Rangplätzen interpretiert werden (vgl. Kühnel und Krebs 2007, 96). Die Angabe der Quartile ist anschaulicher: Das 1. Quartil des Schulabschlusses hat den Wert Hauptschule (Spalte

kumulierte Prozente; Kategorie, in der 25 % der Fälle erreicht werden), der Median hat die Ausprägung „Mittlere Reife“ und das 3. Quartil fällt in die Kategorie „Fachhochschulabschluss“.

6.2.4 Varianz

Im Gegensatz zu Variationsweite und Quartilabstand berücksichtigt die **Varianz** alle Werte einer Verteilung. Sie gibt die **durchschnittliche Variation aller Merkmale** wieder. Die Bezeichnung für die Varianz ist s^2.

- Bei *ungruppierten Daten* wird die Varianz wie folgt berechnet:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\text{SAQ}}{n}\,. \tag{6.8}$$

Die Summe der quadrierten Abweichungen aller Messwerte vom Mittelwert (SAQ) (die nach Punkt 2 in Abschnitt 6.1 minimal ist) wird durch n dividiert. Das Ergebnis – die Varianz – wird deshalb auch als durchschnittliche oder *mittlere quadratische Abweichung* bezeichnet.
Durch die *Quadrierung der Abweichungen* vom Mittelwert werden zwei Dinge erreicht. Zum einen verschwinden die Vorzeichen der Abweichungen. Dies ist auch notwendig, da die durchschnittliche einfache Abweichung aller Messwerte vom arithmetischen Mittel immer null ist, wie wir in Punkt 1 in Abschnitt 6.1 gesehen haben.[1] Zum anderen werden durch die Quadrierung größere Abweichungen vom Mittelwert stärker berücksichtigt als kleine.
Die Summe der quadrierten Abweichungen wird mit Hilfe einer Arbeitstabelle ermittelt (vgl. Tabelle 6.7 auf der folgenden Seite). Die Anzahl der Messwerte n beträgt 11, die durchschnittliche Semesterzahl beträgt $\bar{x} = 13$. Nun kann in der letzten Spalte die Summe der quadrierten Abweichungen vom Mittelwert berechnet werden; sie beträgt SAQ = 74. Im Beispiel ergibt sich also eine Varianz von

$$s^2 = \frac{74}{11} = 6{,}\overline{72} \approx 6{,}73\,.$$

1 Alternativ könnte man die *absoluten Beträge* der einzelnen Abweichungen summieren und durch die Zahl der Beobachtungen dividieren, wodurch man die „AD-Streuung“ erhält. Dieses Maß wird jedoch nur sehr selten verwendet.

Tabelle 6.7: Berechnung der Varianz aus der primären Tafel

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	10	−3	9
2	11	−2	4
3	11	−2	4
4	12	−1	1
5	12	−1	1
6	12	−1	1
7	13	0	0
8	13	0	0
9	14	1	1
10	15	2	4
11	20	7	49
$\sum$	143	0	74

Leider ist diese Zahl schwer zu interpretieren, da durch die Quadrierung die ursprüngliche Maßeinheit (Semester) verloren gegangen ist.

- Bei *gruppierten Daten* werden nicht die einzelnen Merkmalsausprägungen x_i, sondern die Merkmalsausprägungen der Kategorien x_k in die Formel eingebracht. Für jede Kategorie wird die quadrierte Abweichung vom Mittelwert $(x_k - \bar{x})^2$ berechnet und mit ihrer Häufigkeit f_{x_k} multipliziert. Die Berechnung erfolgt wiederum anhand einer Arbeitstabelle (vgl. Tabelle 6.8).

$$s^2 = \frac{\sum_{k=1}^{m} (x_k - \bar{x})^2 \cdot f_{x_k}}{n} \tag{6.9}$$

Im Beispiel ergibt sich:

$$s^2 = \frac{74}{11} = 6{,}\overline{72} \approx 6{,}73\,.$$

Die Varianz beträgt natürlich wieder 6,73.

Tabelle 6.8: Berechnung der Varianz aus den gruppierten Daten

k	x_k	f_{x_k}	$x_k - \bar{x}$	$(x_k - \bar{x})^2$	$(x_k - \bar{x})^2 \cdot f_{x_k}$
1	10	1	−3	9	9
2	11	2	−2	4	8
3	12	3	−1	1	3
4	13	2	0	0	0
5	14	1	1	1	1
6	15	1	2	4	4
7	20	1	7	49	49
$\sum$		11			74

6.2.5 Standardabweichung

Die **Standardabweichung** ergibt sich direkt aus der Quadratwurzel der Varianz. Sie wird mit s bezeichnet.

- Bei ungruppierten Daten lautet die Formel

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\text{SAQ}}{n}}\,. \tag{6.10}$$

- Bei gruppierten Daten lautet die Formel

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{k=1}^{m}(x_k - \bar{x})^2 \cdot f_{x_k}}{n}}\,. \tag{6.11}$$

Im Beispiel erhalten wir

$$s = \sqrt{s^2} = 2{,}59\,.$$

Die Standardabweichung beträgt also 2,59 Semester. Im Gegensatz zur Varianz lässt sich diese Zahl in der ursprünglichen Maßeinheit (hier: Semesterzahl) angeben. Die Standardabweichung ist die *Wurzel aus der mittleren quadratischen Abweichung aller Werte.*

Bei annähernd normalverteilten Merkmalen liegen ca. 68 % aller Werte im Bereich von ± 1 Standardabweichungen um das arithmetische Mittel, ca. 95 % der Werte befinden sich im Bereich von ± 2 Standardabweichungen. Die Körpergröße der in Westdeutschland befragten Männer (ALLBUS 2004, $n = 979$) ist annähernd normalverteilt um ein arithmetisches Mittel von 178 cm mit einer Standardabweichung von 7,3 cm (linke Grafik in Abbildung 6.1, S. 129). Ca. 68 % der befragten westdeutschen Männer sind zwischen 171 $(178 - 7,3)$ und 185 $(178 + 7,3)$ cm groß. Merkmale sind allerdings nur selten normalverteilt. Die Normalverteilung hat jedoch eine große Bedeutung für die schließende Statistik, wie wir in Kapitel 10.3 feststellen werden.

6.2.6 Variationskoeffizient

Merkmale mit einem höheren arithmetischen Mittel weisen häufig auch eine größere Standardabweichung auf. Der *Variationskoeffizient V* relativiert die Standardabweichung am arithmetischen Mittel.

$$V = \frac{s}{\bar{x}} \tag{6.12}$$

Er nimmt einen Wert > 1 an, wenn die Standardabweichung größer ist als das arithmetische Mittel. Multipliziert mit 100 gibt der Variationskoeffizient die Standardabweichung als Prozentwert des arithmetischen Mittelwerts an. Weil sich die Maßeinheit (hier cm) rauskürzt, ist der Variationskoeffizient eine dimensionslose Größe. Er eignet sich deshalb zum Vergleich der Streuung bei zwei Gruppen auch dann, wenn ein Merkmal in unterschiedlichen Maßeinheiten (z. B. Einkommen in US-Dollar und Euro) vorliegt. Für die Semesterzahl beträgt der Variationskoeffizient $2,59/13 = 0{,}20$, also rund 20 % der durchschnittlichen Studiendauer. Die Berechnung des Variationskoeffizienten ist möglich, wenn das betrachtete Merkmal keine negativen Werte annehmen kann, wie es bei ratioskalierten Merkmalen der Fall ist.

In Tabelle 6.9 ist angegeben, ab welchem Skalenniveau die in diesem Kapitel behandelten Lage- und Streuungsmaße sinnvoll interpretiert werden können.

Tabelle 6.9: Univariate Maßzahlen und Skalenniveau

	Skalenniveau			
	nominal	ordinal	intervall	ratio
Modalwert	X	X	X	X
Median		X	X	X
arithmetisches Mittel			X	X
Index qualitativer Variation	X	X	X	X
Quartilabstand		(X)	X	X
Varianz & Standardabweichung			X	X
Variationskoeffizient			(X)	X

Aufgaben zu Lage- und Streuungsparametern

1. Sie haben bei 10 Personen folgende Intelligenzquotienten gemessen:

i	1	2	3	4	5	6	7	8	9	10
x_i	110	160	90	80	111	100	70	100	120	110

   Bitte berechnen Sie die behandelten Lage- und Streuungsparameter und interpretieren Sie diese inhaltlich!
2. In zwei verschiedenen Ländern beträgt das Durchschnittseinkommen $\bar{x} = 1.500$ DM. In Land A beträgt $s = 1.100$ DM in Land B $s = 638$ DM. In welchem Land ist die Einkommensverteilung (bei ansonsten gleichen Bedingungen) gerechter?
3. In der folgenden Tabelle ist die Altersverteilung von Statistik-Kurs-Teilnehmern wiedergegeben (die gleiche Verteilung wurde bereits mit Abbildung 5.5 auf Seite 113 graphisch dargestellt). Bitte berechnen Sie die behandelten Mittel- und Streuungswerte und interpretieren Sie diese inhaltlich!

k	1	2	3	4	5	6	7	8	9	10	11
x_k	19	20	21	22	23	24	25	26	27	28	29
f_{x_k}	1	4	13	13	10	7	6	2	1	1	1

4. Das arithmetische Mittel einer Verteilung beträgt 4, der Median 6. Ist der Modalwert größer als 6, kleiner als beide, oder liegt er zwischen 4 und 6?
5. Sie möchten die Notenverteilung einer Klausur durch einen Mittelwert charakterisieren. Welche(r) Mittelwert(e) ist/sind angemessen und warum?
6. Im Mainzer Mietspiegel sind die mittleren Mieten für jede Wohnungsgruppe anhand des Medians ausgewiesen. Warum?

7 Zusammenhangsmaße

Wenn wir wissen wollen, ob Arbeiter dazu neigen, die SPD zu wählen, ob Vorurteile besonders bei autoritären Persönlichkeiten zu finden sind, oder ein gutes Abitur mit einem guten Studienabschluss einhergeht, dann sind wir auf der Suche nach einem Zusammenhang zwischen zwei Merkmalen.

Zusammenhangsmaße drücken die *Stärke der Beziehung* zwischen zwei Merkmalen aus. Es gibt eine Vielzahl von Zusammenhangsmaßen (Tabelle 7.1). Welches Maß angemessen ist, hängt in erster Linie vom Skalenniveau der Merkmale ab. Da sich für jedes Skalenniveau verschiedene Zusammenhangsmaße berechnen lassen, muss man außerdem berücksichtigen, dass nicht alle Maße zum selben Ergebnis kommen. Darüber hinaus haben alle Zusammenhangsmaße bestimmte Vor- und Nachteile, die bei ihrer Interpretation berücksichtigt werden müssen.

Tabelle 7.1: Zusammenhangsmaße

Merkmal 1	Merkmal 2	Zusammenhangsmaß
		2×2-Tabellen:
dichotom	dichotom	Prozentsatzdifferenz, Odds-Ratio, ϕ (phi), Yules Q
		Mehrfeldertabellen:
nominal	nominal	Cramérs V, C, λ (lambda)
ordinal	ordinal	γ (gamma), tau-Maße (τ_b, τ_c), ρ (rho), Somers d
nominal*	metrisch**	η^2 (eta-Quadrat)
metrisch	metrisch	Kovarianz, Produkt-Moment-Korrelation r

* unabhängiges Merkmal, ** abhängiges Merkmal

Im Folgenden beschränken wir uns auf einige wesentliche Maßzahlen. Eine ausführliche Darstellung findet sich bei Benninghaus (2005). Für das Verständnis multivariater Analyseverfahren ist die Kenntnis von Odds-Ratios,

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_7,

der Kovarianz und der Produkt-Moment-Korrelation zentral. Logarithmierte Odds werden in logistischen Regressionsmodellen als abhängige Variable verwandt. Kovarianz- und Korrelationsmatrizen sind der Ausgangspunkt für Strukturgleichungsmodelle und faktorenanalytische Verfahren (vgl. Reinecke 2005).

7.1 Kreuztabellen und statistische Unabhängigkeit

Angenommen wir möchten herausfinden, ob Ostdeutsche eine andere Einstellung zum Schwangerschaftsabbruch haben als Westdeutsche. Dazu haben wir die Merkmale Einstellung zur Abtreibung (‚Wenn die Frau es will') und Erhebungsgebiet (Befragung in Westdeutschland oder Ostdeutschland) gekreuzt (Tabelle 7.2). Solche Tabellen werden auch als Kontingenztabellen bezeichnet, weil sie die gemeinsame Verteilung zweier Merkmale wiedergeben. Beispielsweise lehnen in Westdeutschland 1401 von 2148 Befragten einen Schwangerschaftsabbruch ab, in Ostdeutschland lehnen 410 von 1086 Befragten einen Schwangerschaftsabbruch ab.

Tabelle 7.2: Einstellung zur Abtreibung nach Erhebungsgebiet (Häufigkeiten)

	Gebiet		
Abtreibung?	West	Ost	Summe
nein	1401	410	1811
ja	747	676	1423
Summe	2148	1086	3234

ALLBUS 2006

Die allgemeine Form einer Kreuztabelle ist in Tabelle 7.3 auf der nächsten Seite dargestellt. Die Variable in den Spalten wird in der Regel mit X bezeichnet, die Variable in den Zeilen mit Y (vgl. Kapitel 5.1.2). Der Laufindex für die Zeilen läuft von $i = 1 \ldots l$, der Laufindex für die Spalten von $j = 1 \ldots m$. In den Zellen stehen die Häufigkeiten f. f_{11} gibt also die Häufigkeit wieder, die sich in der ersten Zeile und ersten Spalte befindet.

Tabelle 7.3: Allgemeine Form einer Kreuztabelle

	x_1	x_2	$\cdots$	x_m	Zeilensumme
y_1	f_{11}	f_{12}	$\cdots$	f_{1m}	$\sum_{j=1}^{m} f_{1j}$
y_2	f_{21}	f_{22}	$\cdots$	f_{2m}	$\sum_{j=1}^{m} f_{2j}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
y_l	f_{l1}	f_{l2}	$\cdots$	f_{lm}	$\sum_{j=1}^{m} f_{lj}$
Spaltensumme	$\sum_{i=1}^{l} f_{i1}$	$\sum_{i=1}^{l} f_{i2}$	$\cdots$	$\sum_{i=1}^{l} f_{im}$	$\sum_{i=1}^{l}\sum_{j=1}^{m} f_{ij} = n$

Die **beobachteten Häufigkeiten** einer Zelle werden allgemein als $f_{b(ij)}$ bezeichnet. f steht auch hier wieder für Häufigkeiten, b gibt an, dass es sich um die beobachteten Häufigkeiten handelt, i kennzeichnet die Zeile und j die Spalte. $f_{b(12)}$ wäre also die beobachtete Häufigkeit der Zelle, die in der ersten Zeile und der zweiten Spalte steht, und dies sind im Beispiel 410 Personen.

Um zu bestimmen wie die Einstellung zum Schwangerschaftsabbruch vom Erhebungsgebiet abhängt, muss spaltenweise prozentuiert werden (Kapitel 5). 65 % der in Westdeutschland Befragten $(1401/2148 \cdot 100)$ lehnen eine Abtreibung ab, jedoch nur 38 % $(410/1086 \cdot 100)$ der ostdeutschen Befragten (Tabelle 7.4, Spaltenprozente in Klammern).

Tabelle 7.4: Beobachtete Häufigkeiten und Spaltenprozente (Kontingenztabelle)

	Gebiet					
Abtreibung?	West		Ost		Gesamt	
nein	1401	(65 %)	410	(38 %)	1811	(56 %)
ja	747	(35 %)	676	(62 %)	1423	(44 %)
Gesamt	2148	(100 %)	1086	(100 %)	3234	(100 %)

ALLBUS 2006

Ost- und westdeutsche Befragte unterscheiden sich offensichtlich in ihren Einstellungen – es besteht ein Zusammenhang zwischen dem Erhebungsgebiet und der Einstellung zur Abtreibung.

Wie würde die Tabelle aussehen, wenn kein Zusammenhang zwischen den Merkmalen besteht, d. h. die beiden Merkmale statistisch *un*abhängig[1] sind? In diesem Fall dürften sich die Einstellungen von ost- und westdeutschen Befragten nicht unterscheiden. Die prozentuale Verteilung der abhängigen Variable (Einstellung zur Abtreibung) wäre dann für jede Ausprägung der unabhängigen Variable (West, Ost) identisch (vgl. Tabelle 7.5). Die Ablehnung eines Schwangerschaftsabbruchs müsste bei ost- *und* westdeutschen Befragten 56 % betragen, die Zustimmung 44 %. Die Häufigkeiten, die dem Modell statistischer Unabhängigkeit entsprechen, werden als *erwartete Häufigkeiten* bezeichnet.

Die **erwarteten Häufigkeiten** lassen sich ganz einfach ermitteln, indem man die Zeilensumme mit der Spaltensumme multipliziert und diesen Wert durch die Gesamtzahl der Befragten (n) dividiert. Die erwarteten Häufigkeiten werden mit $f_{e(ij)}$ bezeichnet.

$$f_{e(ij)} = \frac{\text{Zeilensumme} \cdot \text{Spaltensumme}}{n} \tag{7.1}$$

Im Beispiel ist die erwartete Häufigkeit in der linken oberen Zelle also:

$$f_{e(11)} = \frac{2148 \cdot 1811}{3234} = 1202{,}9\,.$$

Ca. 1203 von 2148 in Westdeutschland befragten Personen müssten einen Schwangerschaftsabbruch ablehnen, wenn kein Zusammenhang zwischen dem Erhebungsgebiet und der Einstellung zum Schwangerschaftsabbruch bestände. Auf die in Gleichung 7.1 beschriebene Art und Weise kann man nun auch die erwarteten Häufigkeiten der anderen Zellen berechnen. Tabellen, die die erwarteten Häufigkeiten beinhalten, werden auch als Indifferenztabellen bezeichnet.

1 Das Konzept der statistischen Unabhängigkeit bezieht sich eigentlich auf eine Grundgesamtheit (vgl. Kapitel 12). In diesem Kapitel begnügen wir uns mit Aussagen über Stichprobendaten.

Tabelle 7.5: Erwartete Häufigkeiten und Spaltenprozente bei statistischer Unabhängigkeit (Indifferenztabelle)

	Gebiet					
Abtreibung?	West		Ost		Gesamt	
nein	1202,9	(56%)	608,2	(56%)	1811	(56%)
ja	945,2	(44%)	477,9	(44%)	1423	(44%)
Gesamt	2148	(100%)	1086	(100%)	3234	(100%)

ALLBUS 2006

Statistische Unabhängigkeit ist eine symmetrische Eigenschaft. Wenn die prozentuale Verteilung innerhalb der Spalten identisch ist, dann ist auch die prozentuale Verteilung innerhalb der Zeilen identisch, wie man durch Zeilenprozentuierung der erwarteten Häufigkeiten leicht feststellen kann.

7.2 Maße für zwei dichotome Merkmale

7.2.1 Prozentsatzdifferenz

Ein einfach zu interpretierendes Maß für den Zusammenhang in 2×2-Tabellen ist die Prozentsatzdifferenz. Die Prozentsatzdifferenz der Ablehnung eines Schwangerschaftsabbruchs zwischen West- und Ostdeutschen lässt sich aus den beobachteten Häufigkeiten (Tabelle 7.4) berechnen,

$$\left(\frac{1401}{2148} - \frac{410}{1086}\right) \cdot 100 = 27 \quad \text{Prozentpunkte,} \tag{7.2}$$

oder kann direkt aus den Spaltenprozenten (Tabelle 7.4) bestimmt werden: $65\,\% - 38\,\% = 27$ Prozentpunkte. Die Prozentsatzdifferenz hat einen Wertebereich von –100 bis +100. Je größer der absolute Wert, umso stärker der Zusammenhang. Bei statistischer Unabhängigkeit zweier Merkmale (Tabelle 7.5) ist die Prozentsatzdifferenz null.

Die Prozentsatzdifferenz ist ein asymmetrisches Zusammenhangsmaß. Sie hängt davon ab, welches Merkmal als abhängig und welches als unabhängig betrachtet wird. Hier wurde die Einstellung zum Schwangerschafts-

abbruch in Abhängigkeit vom Erhebungsgebiet betrachtet. Wird dagegen das Erhebungsgebiet als abhängiges Merkmal betrachtet (was wenig Sinn ergibt, da wohl niemand wegen seiner Einstellung zur Abtreibung den Wohnort verlagert), dann ändert sich die Prozentsatzdifferenz. Nur $(410/1811) \cdot 100 = 22,6\,\%$ der Gegner eines Schwangerschaftsabbruchs leben in Ostdeutschland, aber $(676/1423) \cdot 100 = 47,5\,\%$ der Befürworter. Die Prozentsatzdifferenz beläuft sich auf $[(410/1811) - (676/1423)] \cdot 100 = -25$ Prozentpunkte.

7.2.2 Odds-Ratio

Für das Verständnis logistischer Regressionen sind Odds (Chancen) und Odds-Ratios (Chancenverhältnisse) von zentraler Bedeutung. Für eine binäre abhängige Variable geben die *Odds* die Häufigkeit des interessierenden Ereignisses (Ablehnung eines Schwangerschaftsabbruchs) zur Häufigkeit des Gegenereignisses (Befürwortung eines Schwangerschaftsabbruchs) an.

$$\text{Odds} = \frac{\text{Häufigkeit Ereignis}}{\text{Häufigkeit Gegenereignis}} \qquad (7.3)$$

Treten beide Kategorien gleich häufig auf, dann betragen die Odds 1. Die Odds haben einen Wert größer eins, wenn das interessierende Ereignis häufiger auftritt als das Gegenereignis. Sie sind kleiner 1, wenn das interessierende Ereignis seltener auftritt als das Gegenereignis. Odds haben einen Wertebereich von 0 bis $+\infty$. Für die westdeutschen Befragten betragen die Odds der Ablehnung eines Schwangerschaftsabbruchs

$$\text{Odds}_{\text{West}} = \frac{1401}{747} = 1,88\,. \qquad (7.4)$$

Die Ablehnung des Schwangerschaftsabbruchs ist für westdeutsche Befragte 1,9-mal häufiger als die Zustimmung. Es kommen 1,9 Ablehnungen auf eine Zustimmung. Umgekehrt ist das Verhältnis von Zustimmung zu Ablehnung 1 zu 1,9 (1/1,9=0,53). Für ostdeutsche Befragte betragen die Odds

$$\text{Odds}_{\text{Ost}} = \frac{410}{676} = 0,61\,. \tag{7.5}$$

Für Ostdeutsche ist die Ablehnung eines Schwangerschaftsabbruchs seltener als die Befürwortung (Odds < 1). Die Chance der Ablehnung eines Schwangerschaftsabbruchs beträgt 0,6 zu 1. Die Chancen einer Befürwortung $1/0,6 = 1,\overline{66}$. Die Odds für westdeutsche und ostdeutsche Befragte werden als konditionale Odds bezeichnet. Konditionale Odds sind durch die Ausprägungen (West, Ost) eines unabhängigen Merkmals (Erhebungsgebiet) bedingt.

Odds sind keine Wahrscheinlichkeiten. Die über den Anteil geschätzte Wahrscheinlichkeit der Ablehnung eines Schwangerschaftsabbruchs bei Ostdeutschen beträgt 0,38. Die Wahrscheinlichkeit setzt die interessierende Kategorie allen Kategorien ins Verhältnis, die Odds setzen dagegen die interessierende Kategorie zur Gegenkategorie ins Verhältnis. Wahrscheinlichkeiten können leicht in Odds umgerechnet werden. Dazu wird die Wahrscheinlichkeit des Ereignisses p durch die Wahrscheinlichkeit aller anderen Ereignisse $(1 - p)$ dividiert (Gleichung 7.6). Für Ostdeutsche also $0,38/(1 - 0{,}38) = 0{,}61$.

$$\text{Odds} = \frac{p}{1 - p} \tag{7.6}$$

Besteht ein Zusammenhang zwischen dem Erhebungsgebiet und der Einstellung zur Abtreibung, dann unterscheiden sich die Odds der westdeutschen Befragten von den Odds der ostdeutschen Befragten. Um die konditionalen Odds miteinander zu vergleichen, wird das Verhältnis, die Odds-Ratio (das Chancenverhältnis), gebildet.

$$\text{Odds-Ratio} = \frac{\text{Odds}_1}{\text{Odds}_2} \tag{7.7}$$

Die Odds-Ratio beträgt für west- und ostdeutsche Befragte $1,88/0,61 = 3,1$. Die Chance der Westdeutschen einen Schwangerschaftsabbruch abzulehnen, beträgt das 3,1fache der Chance der Ostdeutschen. Oder um-

gekehrt ausgedrückt: Die Chance der Ostdeutschen, Schwangerschaftsabbrüche abzulehnen, beträgt ca. ein Drittel $(1/3,1 = 0,32)$ der Chance für Westdeutsche.

Odds-Ratios haben einen Wertebereich von 0 bis ∞. Sind die beiden konditionalen Odds identisch (kein Zusammenhang), dann nimmt die Odds-Ratio den Wert 1 an. Ein Odds-Ratio > 1 bedeutet, dass die Odds für Gruppe 1 größer sind als die Odds für Gruppe 2. Ein Odds-Ratio < 1 zeigt, dass die Odds für Gruppe 1 kleiner sind als für Gruppe 2. Je weiter der Wert von 1 entfernt ist, umso stärker ist der Zusammenhang zwischen den beiden Merkmalen. Ein Odds-Ratio von 4 gibt einen stärkeren Zusammenhang wieder als ein Odds-Ratio von 1,5. Ein Odds-Ratio von 0,25 drückt einen stärkeren Zusammenhang aus als ein Odds-Ratio von 0,4. Um die Stärke des Zusammenhangs von Odds-Ratios $<$ 1 mit Odds-Ratios $>$ 1 zu vergleichen, wird der Kehrwert der Odds-Ratios betrachtet, die kleiner 1 sind. Eine Odds-Ratio von 2 drückt einen gleich starken Zusammenhang aus wie eine Odds-Ratio von 0,5 (1/0,5=2), allerdings in unterschiedlicher Richtung. Dies wird deutlich, wenn man den natürlichen Logarithmus der Odds-Ratios berechnet: $\ln 0,5 = -0,69$ und $\ln 2 = +0,69$.

Odds-Ratios zählen zu den symmetrischen Zusammenhangsmaßen. Die Größe der Odds-Ratio hängt also nicht davon ab, welches der beiden Merkmale als abhängig oder unabhängig betrachtet wird.

Ein Kritikpunkt an Odds-Ratios besteht darin, dass sie nichts mehr über die Größe der Odds (und Wahrscheinlichkeiten) aussagen: Eine Odds-Ratio von beispielsweise 2 kann daraus resultieren, dass die Odds für Gruppe 1 0,02 und für Gruppe 2 0,01 betragen (0,02/0,01=2). Die Odds für das interessierende Ereignis sind dann für Gruppe 1 zwar doppelt so hoch, allerdings auf einem sehr niedrigen Niveau.

7.3 Maße für zwei nominalskalierte Merkmale

Prozentsatzdifferenz und Odds-Ratio sind Maße für die Stärke des Zusammenhangs in 2×2-Tabellen. In Mehrfeldertabellen lassen sich mehrere Prozentsatzdifferenzen und Odds-Ratios berechnen. Auch für Mehrfeldertabellen gibt es Maßzahlen, die den Zusammenhang zwischen zwei Merkmalen in einer *einzigen* Zahl ausdrücken. Nominalskalierte Zusammenhangsmaße haben einen Wertebereich von 0 bis 1. Null bedeutet kein Zusammenhang; eine eins gibt ein perfekten Zusammenhang an. Sie sind

vorzeichenlos, weil die Ausprägungen nominalskalierter Merkmale keine Rangordnung aufweisen.

7.3.1 Kontingenzkoeffizient C und Cramérs V

Cramérs V und der Kontingenzkoeffizient C sind χ^2-basierte Zusammenhangsmaße. Ausgangspunkt zur Berechnung von χ^2 (chi^2) sind die beobachteten Häufigkeiten $f_{b(ij)}$ und die bei statistischer Unabhängigkeit erwarteten Häufigkeiten $f_{e(ij)}$, die in den Tabellen 7.4 und 7.5 auf S. 143 dargestellt sind.

Je größer die Differenz zwischen beobachteten und erwarteten Häufigkeiten, umso stärker weichen die Daten vom Modell statistischer Unabhängigkeit ab. Wie man in Tabelle 7.4 sieht, lehnen 1401 westdeutsche Befragte einen Schwangerschaftsabbruch ab (linke obere Zelle). Bei statistischer Unabhängigkeit der beiden Merkmale Erhebungsgebiet und Einstellung zum Schwangerschaftsabbruch müssten ca. 1203 (1202,9) westdeutsche Befragte einen Schwangerschaftsabbruch ablehnen (Tabelle 7.5). Die Differenz zwischen den beobachteten und den erwarteten Häufigkeiten beträgt für die linke obere Zelle $f_{b(ij)} - f_{e(ij)} = 1401 - 1202,9 = 198,1$. Es haben mehr westdeutsche Befragte (198,1) einen Schwangerschaftsabbruch abgelehnt, als wir es bei statistischer Unabhängigkeit beider Merkmale erwarten würden.

Die Differenz zwischen beobachteten und erwarteten Häufigkeiten muss für jede Zelle berechnet werden. Die Summe dieser einfachen Abweichungen für alle Zellen ist bei *jeder* Kreuztabelle null und deshalb als Maß der Abhängigkeit beider Merkmale ungeeignet. Die Differenz zwischen erwarteten und beobachteten Häufigkeiten wird deshalb *quadriert*: $(f_{b(ij)} - f_{e(ij)})^2$. Durch die Quadrierung fallen die negativen Vorzeichen weg. Zudem werden große Abweichungen der beobachteten von den erwarteten Häufigkeiten stärker gewichtet als kleinen Abweichungen. Ob eine bestimmte Abweichung als groß oder klein zu bewerten ist, hängt außerdem davon ab, wie groß die erwartete Häufigkeit ist: Sowohl in der Zelle links oben (f_{11}) als auch in der Zelle rechts unten (f_{22}) beträgt die Differenz zwischen erwarteten und beobachteten Häufigkeiten 198,1. Diese Differenz fällt bei einer erwarteten Häufigkeit von 1202,9 (links oben) weniger stark ins Gewicht als bei einer erwarteten Häufigkeit von 477,9 (rechts unten). Die quadrierte Differenz $(f_{b(ij)} - f_{e(ij)})^2$ wird deshalb an der erwarteten Häufigkeit einer Zelle $f_{e(ij)}$ relativiert:

$$\frac{(f_{b(ij)} - f_{e(ij)})^2}{f_{e(ij)}} \, .$$

Die Maßzahl χ^2 ist nun nichts anderes als die Summe dieser quadrierten und an den erwarteten Häufigkeiten relativierten Abweichungen für alle Zellen:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(f_{b(ij)} - f_{e(ij)})^2}{f_{e(ij)}} \, . \tag{7.8}$$

Die Summenzeichen geben an, dass $(f_{b(ij)} - f_{e(ij)})/f_{e(ij)}$ für alle Zellen berechnet und dann addiert wird. Für den Zusammenhang zwischen Erhebungsgebiet und Einstellung zur Abtreibung resultiert ein χ^2-Wert von

$$\begin{aligned}\chi^2 =& \frac{(1401 - 1202,9)^2}{1202,9} + \frac{(410 - 608,2)^2}{608,2} + \frac{(747 - 945,2)^2}{945,2} + \frac{(676 - 477,9)^2}{477,9} \\ =& 221.\end{aligned}$$

Ein χ^2-Wert von null bedeutet kein Zusammenhang. χ^2 ist als Maß der Stärke eines Zusammenhangs jedoch ungeeignet, weil der χ^2-Wert von der Zahl der Beobachtungen abhängt. Würde man die beobachteten Häufigkeiten in den Zellen von Tabelle 7.4 verdoppeln, dann würde sich auch der χ^2-Wert verdoppeln, ohne dass sich an der prozentualen Verteilung – dem Zusammenhang – etwas ändert. (Wir werden in Kapitel 12 jedoch sehen, dass χ^2 für die schließende Statistik eine große Bedeutung hat.)

Aus diesem Grund wurden Maße vorgeschlagen, die χ^2 normieren: Der Kontingenzkoeffizient C und Cramérs V.

- Der Kontingenzkoeffizient C hat einen Wertebereich zwischen 0 und einem definierten Maximum C_{max}. C berechnet sich nach der Formel

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \, . \tag{7.9}$$

n ist hier wieder die Anzahl der Messwerte. Die maximale Größe des Kontingenzkoeffizienten ist abhängig von der Zahl der Zeilen bzw. Spalten einer Tabelle und läßt sich nach

$$C_{max} = \sqrt{\frac{R-1}{R}}; \qquad R = min(l,m) \tag{7.10}$$

bestimmen. R ist das Minimum der Zeilen- bzw. Spaltenzahl. In einer 2×2-Tabelle ist $R = 2$, in einer Tabelle mit 3 Zeilen und 4 Spalten ist R=3, in einer Tabelle mit 4 Zeilen und 3 Spalten ist R ebenfalls 3. In einer 2×2-Tabelle kann C maximal den Wert $C_{max} = \sqrt{(2-1)/2} =$ 0,707 annehmen.
Im Beispiel resultiert ein Zusammenhang der Stärke

$$C = \sqrt{\frac{221}{221 + 3234}} = 0{,}25\,.$$

- Weil Kontingenzkoeffizienten aus Tabellen unterschiedlicher Größe nur bedingt vergleichbar sind, ist Cramérs V vorzuziehen. Cramérs V normiert den χ^2-Wert, indem dieser durch den maximal erreichbaren χ^2-Wert in einer Tabelle dividiert wird (Gleichung 7.11). In einer 2×2-Tabelle kann χ^2 maximal so groß sein wie die Zahl der Beobachtungen n, hier also 3234. In Mehrfeldertabellen ist $\chi^2_{max} = n(R-1)$, wobei R auch hier dem Minimum der Zeilen- bzw. Spaltenzahl entspricht.

$$\text{Cramérs V} = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{n \cdot (R-1)}} \tag{7.11}$$

Cramérs V nimmt für unser Beispiel den Wert

$$\text{Cramérs V} = \sqrt{\frac{221}{3234 \cdot (2-1)}} = 0,26$$

an. Für 2×2-Tabellen ist Cramérs V vom Betrag identisch mit ϕ (vgl. zur Herleitung Kühnel und Krebs 2007, 336). Aufgrund unterschiedlicher Berechnungsarten kann ϕ allerdings negative Vorzeichen annehmen.

Cramérs V und der Kontingenzkoeffizient fallen also ungefähr gleich hoch aus. Ein Cramérs V von 0,26 liegt im unteren Bereich auf einer Skala von 0 bis 1, deutet also auf einen geringen Zusammenhang hin. Zu beachten ist allerdings, dass die empirisch zu beobachtenden Werte von Cramérs V in der Regel weit vom Maximalwert 1 entfernt sind. Cramérs V-Werte zwischen 0,1 und 0,2 werden als schwache Zusammenhänge angesehen. Die Interpretation der Stärke des Zusammenhangs ist mit Ausnahme der Extremwerte 0 – kein Zusammenhang – und 1 – perfekter Zusammenhang – jedoch nicht eindeutig. In der Literatur herrschen unterschiedliche Auffassungen darüber, ab wann ein durch Cramérs V gemessener Zusammenhang als stark zu bezeichnen ist.

Ein anderer Kritikpunkt ist der, dass eine einzige Maßzahl nur wenig über die Art des Zusammenhangs zwischen zwei mehrfach gestuften, nominalskalierten Merkmalen aussagt. Für den Zusammenhang zwischen der Religionszugehörigkeit und der Einstellung zur Abtreibung (Tabelle 7.6) beträgt Cramérs V 0,17 ($\chi^2 = 61{,}99$). Dieser Wert deutet auf einen schwachen Zusammenhang hin. Wie sich die Angehörigen verschiedener Religionen/Konfessionen in ihren Einstellungen unterscheiden, geht daraus nicht hervor. Dazu muss das Antwortverhalten der einzelnen Religionsgruppen miteinander verglichen werden – am einfachsten anhand der prozentualen Verteilung. Man sieht, dass katholische Befragte und Befragte einer nichtchristlichen Glaubensgemeinschaft einen Schwangerschaftsabbruch prozentual am stärksten ablehnen. Die geringste Ablehnung äußern konfessionslose Befragte (Spalte ‚Keine').

Tabelle 7.6: Einstellung zur Abtreibung nach Religion - Beobachtete Häufigkeiten und Spaltenprozente

Abtreibung?	Religion											
	Evang./ Freik.		Kath.		Andere christl.		Nicht christl.		Keine		Gesamt	
nein	541	(64%)	564	(71%)	33	(66%)	78	(80%)	177	(50%)	1393	(65%)
ja	301	(36%)	226	(29%)	17	(34%)	19	(20%)	180	(50%)	743	(35%)
Gesamt	842	(100%)	790	(100%)	50	(100%)	97	(100%)	357	(100%)	2136	(100%)

ALLBUS 2006. Westdeutsche Befragte.

7.3.2 Das PRE-Maß lambda (λ)

Die Stärke des Zusammenhangs zwischen zwei Merkmalen wird bei den **PRE-Maßen** (PRE = Proportional Reduction in Error) daran gemessen, wie gut die Werte eines abhängigen Merkmals durch die Kenntnis eines unabhängigen Merkmals vorhergesagt werden können. λ (lambda) ist ein PRE-Maß für nominale Daten.

Zur Illustration wird die Bundestagswahl 1994 herangezogen. Gegen den seit 1982 amtierenden Kanzler Helmut Kohl trat 1994 Rudolf Scharping als Spitzenkandidat für die SPD an, nachdem es Oskar Lafontaine 1990 nicht gelungen war, einen Regierungswechsel herbeizuführen. Vor der Wahl erhob die Forschungsgruppe Wahlen die Wahlabsicht in einer Umfrage. Von 881 westdeutschen Befragten, die eine Partei angaben, wollten 350 Befragte die CDU/CSU und 345 Befragte die SPD wählen. 186 Befragte wollten ihre Stimme einer anderen Partei zukommen lassen.

Bei der Berechnung eines PRE-Maßes wird zunächst versucht, *den Wert der abhängigen Variablen ohne Kenntnis einer unabhängigen Variablen zu prognostizieren* (Schritt 1). Die beste Prognose für ein *nominales* Merkmal ist deren Modalwert (Modus). Bei Prognose des Modus werden die wenigsten Fehler gemacht. Der Modus der Wahlabsicht ist die CDU/CSU. Bei Prognose einer CDU/CSU-Wahlabsicht liegen wir in 350 Fällen richtig, denn so viele Personen haben ja tatsächlich eine Wahlabsicht zugunsten der CDU/CSU angegeben. In 531 (345 + 186) Fällen – das sind die Befragten, die nicht CDU/CSU wählen wollten – irren wir uns. Bei der Vorhersage einer anderen Partei, z. B. der SPD, würden wir jedoch in noch mehr Fällen – nämlich bei 536 (350 + 186) Personen – falsch liegen. Die Summe der Fehler bei Prognose der abhängigen Variablen ohne Berücksichtigung einer unabhängigen Variablen nennen wir Fehler_1.

Zur Prognose der abhängigen Variable wird nun eine unabhängige Variable herangezogen (Schritt 2). Zur Prognose der Wahlabsicht verwenden wir die Kanzlerpräferenz. In Abbildung 7.7 auf der nächsten Seite ist der Zusammenhang zwischen der Kanzlerpräferenz und der Wahlabsicht wiedergegeben. Für jede Ausprägung der unabhängigen Variablen wird der Wert der abhängigen Variablen nun getrennt prognostiziert. Der beste Prognosewert ist der Modalwert der konditionalen, d. h. durch die Ausprägung des unabhängigen Merkmals bedingten, Verteilung des abhängigen Merkmals. Für die 444 Befragten, die Helmut Kohl als Kanzler bevorzugen,

prognostizieren wir die Wahl der CDU/CSU, weil die CDU/CSU hier am häufigsten genannt wurde. In 335 Fällen treffen wir mit dieser Prognose ins Schwarze, in 109 Fällen (25 SPD-Wähler und 84 Wähler anderer Parteien) irren wir uns. Unsere Prognose für die 437 Befragten, die Rudolf Scharping bevorzugen, lautet dagegen SPD (Modalkategorie). Hier schätzen wir die Wahlabsicht von 320 Personen richtig ein, dagegen irren wir uns in 117 (15 + 102) Fällen – denjenigen Befragten, die trotz einer Präferenz für Rudolf Scharping nicht die SPD wählen wollen. Die Summe der Fehler, die wir trotz Berücksichtigung der unabhängigen Variablen begehen, nennen wir Fehler$_2$. Allgemein berechnet man Fehler$_2$, indem man *für jede Ausprägung der unabhängigen Variablen die Prognosefehler berechnet und diese summiert.* Im Beispiel berechnen wir also für die erste Ausprägung der unabhängigen Variablen (Helmut Kohl als bevorzugter Kanzlerkandidat) 109 Fehler und für die zweite Ausprägung (Rudolf Scharping als bevorzugter Kanzlerkandidat) 117 Fehler. Fehler$_2$ beträgt also $109 + 117 = 226$.

Tabelle 7.7: Zusammenhang von Kanzlerpräferenz und Wahlabsicht

	Kanzlerpräferenz		
Wahlabsicht	Kohl	Scharping	Summe
CDU/CSU	335	15	350
SPD	25	320	345
Andere	84	102	186
Summe	444	437	881

Quelle: Forschungsgruppe Wahlen, Blitzumfrage Oktober 1994, nur Westdeutsche

Der letzte Schritt besteht nun in der *Ermittlung des PRE-Maßes* (Schritt 3):

$$\text{PRE-Maß} = \frac{(\text{Fehler}_1 - \text{Fehler}_2)}{\text{Fehler}_1} . \qquad (7.12)$$

Diese Formel ist für alle PRE-Maße identisch. Lediglich die Berechnung der Fehler unterscheidet sich, wie wir bei η^2 sehen werden. Je kleiner Fehler$_2$ im Vergleich zu Fehler$_1$ ist, umso besser wird die abhängige durch die unabhängige Variable prognostiziert. Zusätzlich wird die Differenz zwischen Fehler$_1$ und Fehler$_2$ auf einen Wertebereich zwischen 0 und 1 normiert, indem durch Fehler$_1$ dividiert wird.

λ kann Werte zwischen 0 und +1 annehmen. λ ist 0, wenn die unabhängige Variable die Prognose nicht verbessert und den Wert 1, wenn wir den Wert der abhängigen Variable in *allen* Fällen durch die unabhängige Variable richtig vorhersagen. λ kann – multipliziert mit 100 – wie jedes PRE-Maß anschaulich prozentual interpretiert werden.

Inwieweit wurde die Vorhersage der Wahlabsicht nun durch die Kenntnis des bevorzugten Kanzlers verbessert? Im Beispiel ergibt sich:

$$\lambda = \frac{(\text{Fehler}_1 - \text{Fehler}_2)}{\text{Fehler}_1} = \frac{(531 - 226)}{531} = 0{,}57\,.$$

Die Fehler bei der Prognose der Wahlabsicht werden durch die Kenntnis der Kanzlerpräferenz also um 57 % verringert.

λ ist ein *asymmetrisches Maß*. Je nachdem, welche Variable als abhängig und welche als unabhängig betrachtet wird, ergibt sich also ein unterschiedlicher Wert für λ. Für die Vorhersage des präferierten Kanzlerkandidaten durch die Wahlabsicht (was eine Vertauschung der abhängigen und der unabhängigen Variablen bedeutet) ergibt sich ein λ von 0,72. Die unterschiedlichen Werte resultieren aus der unterschiedlichen Berechnung von Fehler_1: einmal liegt dessen Berechnung die Wahlabsicht zugrunde, das andere Mal die Kanzlerpräferenz.

Für den Zusammenhang zwischen Erhebungsgebiet und Einstellung zur Abtreibung (Tabelle 7.4) beträgt λ 0,19. Der durch λ gemessene Einfluss der Religionszugehörigkeit auf die Einstellung zum Schwangerschaftsabbruch (Tabelle 7.6) liegt nahe null, nämlich bei 0,04. Cramérs V deutet dagegen auf einen schwachen Zusammenhang hin.

An diesem Beispiel lässt sich ein Nachteil von λ demonstrieren. λ kann einen Wert von null annehmen, obwohl andere Zusammenhangsmaße wie Cramérs V einen Zusammenhang ausweisen. Und zwar dann, wenn die Modalkategorie des abhängigen Merkmals für jede Ausprägung des unabhängigen Merkmals identisch ist. In Tabelle 7.6 prognostizieren wir für jede Konfessions-/Religionszugehörigkeit die Ausprägung ‚nein'. Lediglich für Konfessionslose (‚Keine') prognostizieren wir ‚ja'. Selbst bei Konfessionslosen sind beide Ausprägungen des abhängigen Merkmals (nein/ja) jedoch ungefähr gleich stark besetzt, so dass sich die Prognose durch Kenntnis der Religionszugehörigkeit nur geringfügig verbessert.

7.4 Maße für zwei ordinalskalierte Merkmale

Bei ordinalskalierten Merkmalen können wir neben der Stärke auch die Richtung des Zusammenhangs angeben. Ein positiver (negativer) Zusammenhang liegt vor, wenn höhere Werte auf der einen Variablen mit höheren (niedrigeren) Werten auf der anderen Variablen einhergehen. Zusammenhangsmaße für ordinalskalierte Merkmale haben einen Wertebereich von -1 bis $+1$, wobei -1 ein perfekter negativer Zusammenhang ist, $+1$ dagegen ein perfekter positiver Zusammenhang. Null bedeutet wie bei den nominalskalierten Merkmalen, dass kein Zusammenhang vorliegt.

Es gibt eine Vielzahl an Zusammenhangsmaßen für ordinalskalierte Merkmale. Die am häufigsten verwendeten sind Maße, die auf dem Paarvergleich basieren: Kendalls tau-Maße (τ_b und τ_c) und gamma (γ). Wir beschränken uns auf γ. γ hat den Vorteil, dass es wie ein PRE-Maß interpretiert werden kann.

gamma (γ)

Zum Verständnis der Berechnung von γ ist es notwendig, sich die Logik des Paarvergleichs vor Augen zu führen. Als Beispiel dient der Zusammenhang zwischen dem Schulabschluss und dem politischen Interesse (Tabelle 7.8). Beide Variablen sind ordinalskaliert, d. h. sie weisen eine Ordnung auf. Wir unterstellen, dass der Schulabschluss das politische Interesse beeinflusst.

Tabelle 7.8: Kreuztabelle zwischen Bildung und politischem Interesse

	Schulabschluss			
Pol. Interesse	Hauptschule	Realschule	FH/Abitur	Summe
Kein	228	72	10	310
Wenig	386	209	67	662
Mittel	741	460	244	1445
Stark	219	189	229	637
Sehr stark	75	87	103	265
Summe	1649	1017	653	3319

Quelle: ALLBUS 1994

Ein Befragter kann z. B. einen „Hauptschulabschluss“ und „kein politisches Interesse“ haben, ein anderer einen „Realschulabschluss“ und „wenig poli-

tisches Interesse". In der Logik des Paarvergleichs wird dieses Paar als **konkordant** oder **gleichgerichtet** bezeichnet, da der zweite Befragte einen höheren Schulabschluss *und* ein höheres politisches Interesse hat als der erste Befragte. Ein Paar wird als konkordant bezeichnet, wenn die Person, die einen höheren Wert auf der einen Variablen aufweist, auch einen höheren Wert auf der anderen Variablen hat. Konkordante Paare deuten auf einen *positiven Zusammenhang* zwischen zwei Variablen hin. Insgesamt gibt es 228 Personen mit „Hauptschulabschluss" und „keinem politischen Interesse"; 209 Personen haben einen „Realschulabschluss" und „wenig politisches Interesse". Alle 209 Personen dieser Zelle haben einen höheren Schulabschluss *und* ein größeres politisches Interesse als die 228 Personen, die in der Zelle links oben verortet sind, d. h. sie weisen bei *beiden* Merkmalen „mehr" auf. Die Anzahl konkordanter Paare (N_c) dieser beiden Zellen berechnet sich aus der Multiplikation der Zellhäufigkeiten, also $228 \times 209 = 47.652$ Paare, denn jeder Befragte aus der einen Zelle bildet mit jedem Befragten der anderen Zelle ein Paar. Alle Personen, die sich in Zellen rechts *und* unterhalb einer Ausgangszelle befinden, haben auf beiden Merkmalen einen höheren Wert. Die Zahl der konkordanten Paare für die linke obere Zelle (228 Befragte) berechnet sich daher als $228 \cdot (209 + 67 + 460 + 244 + 189 + 229 + 87 + 103)$.

Zur Ermittlung der *Gesamtzahl konkordanter Paare* wird jede Zelle der Tabelle einmal zur Ausgangszelle. Zu den Zellen in der untersten Zeile sowie in der äußersten rechten Spalte existieren keine Zellen, die rechts *und* unterhalb liegen. Man fängt am besten in der linken oberen Zelle mit der Berechnung an:

$$
\begin{aligned}
N_c = {} & 228 \cdot (209 + 460 + 189 + 87 + 67 + 244 + 229 + 103) \\
& + 72 \cdot (67 + 244 + 229 + 103) \\
& + 386 \cdot (460 + 189 + 87 + 244 + 229 + 103) \\
& + 209 \cdot (244 + 229 + 103) \\
& + 741 \cdot (189 + 87 + 229 + 103) \\
& + 460 \cdot (229 + 103) \\
& + 219 \cdot (87 + 103) \\
& + 189 \cdot (103) \\
= {} & 1699501
\end{aligned}
$$

Es kann jedoch vorkommen, dass eine Person einen „Realschulabschluss" erworben hat und nur „wenig" politisch interessiert ist, eine andere Person

dagegen einen „Hauptschulabschluss“ und ein „starkes politisches Interesse“ aufweist. Ein solches Paar wird **diskordant** oder **ungleichgerichtet** genannt, da die zweite Person gegenüber der ersten bei der einen Variablen „weniger“ aufweist, bei der anderen Variablen dagegen „mehr“. Diskordante Paare geben einen *negativen Zusammenhang* zwischen zwei Variablen wieder, da höhere Werte auf der einen Variable mit niedrigeren Werten auf der anderen Variablen einhergehen. Auch für die *Gesamtzahl diskordanter Paare* (N_d) gibt es eine allgemeine Berechnungsvorschrift: Alle Häufigkeiten in Zellen, die *links und unterhalb* einer Ausgangszelle liegen, werden mit der Häufigkeit der Ausgangszelle multipliziert, wobei auch hier jede Zelle einmal zur Ausgangszelle wird. Zu Zellen in der ganz linken Spalte und der untersten Zeile existieren keine Zellen, die links und unterhalb liegen – hier kann es also keine diskordanten Paare geben. Mit der Berechnung starten wir in der Zelle rechts oben:

$$\begin{aligned} N_d = {} & 10 \cdot (386 + 741 + 219 + 75 + 209 + 460 + 189 + 87) \\ & + 72 \cdot (386 + 741 + 219 + 75) \\ & + 67 \cdot (741 + 219 + 75 + 460 + 189 + 87) \\ & + 209 \cdot (741 + 219 + 75) \\ & + 244 \cdot (219 + 75 + 189 + 87) \\ & + 460 \cdot (219 + 75) \\ & + 229 \cdot (75 + 87) \\ & + 189 \cdot (75) \\ = {} & 786537 \end{aligned}$$

In unserem Beispiel ermitteln wir also 1.699.501 konkordante und 786.537 diskordante Paare. Überwiegen in einer Tabelle – wie in diesem Fall – die konkordanten Paare, so liegt ein *positiver Zusammenhang* vor. Der Zusammenhang zwischen zwei Variablen ist *negativ*, wenn es mehr diskordante als konkordante Paare gibt. Zwischen beiden Variablen besteht *kein Zusammenhang*, wenn die Zahl konkordanter und diskordanter Paare gleich groß ist.

Bei der Berechnung des Ordinalmaßes γ wird nun einfach die Differenz zwischen konkordanten und diskordanten Paaren ins Verhältnis zu allen konkordanten und diskordanten Paaren gesetzt:

$$\gamma = \frac{N_c - N_d}{N_c + N_d}\,. \qquad (7.13)$$

γ nimmt Werte zwischen -1 und $+1$ an. Das Vorzeichen gibt an, ob ein negativer oder positiver Zusammenhang vorliegt. Je größer der Unterschied zwischen der Zahl konkordanter und diskordanter Paare, umso stärker ist der Zusammenhang und damit der Betrag von γ. γ erreicht sein Maximum von $+1$, wenn in einer Tabelle nur konkordante, jedoch keine diskordanten Paare vorliegen. Den Wert -1 nimmt γ nur dann an, wenn es in einer Tabelle nur diskordante, aber keine konkordanten Paare gibt. γ ist ein symmetrisches Maß. Der Wert von γ ist also unabhängig davon, welche der Variablen als abhängig bzw. unabhängig betrachtet wird.

Zwischen Schulabschluss und politischem Interesse ermitteln wir einen Wert von

$$\gamma = \frac{1699501 - 786537}{1699501 + 786537} = \frac{912964}{2486038} = 0{,}367\,.$$

Ein Wert von 0,367 deutet auf einen relativ starken positiven Zusammenhang hin. Das heißt: Je höher der Bildungsabschluss, umso stärker ist das politische Interesse. Ein negatives Vorzeichen würde bedeuten, dass mit höherer Bildung das politische Interesse abnimmt. Bei der Interpretation des Vorzeichens ist allerdings die Kodierung der Variablen zu beachten. Die Berechnung des Kennwertes erfolgt ja nur anhand der zugewiesenen Zahlenwerte, ungeachtet der dahinterstehenden inhaltlichen Merkmalsausprägungen.[2]

γ ist ebenfalls ein PRE-Maß. $|\gamma|$ kann – multipliziert mit 100 – als proportionale Fehlerreduktion interpretiert werden. Im Beispiel wird der Vorhersagefehler um 36,7 % verringert, wenn zur Prognose die Schulbildung der Befragten berücksichtigt wird. Ein γ von $-,50$ würde bedeuten, dass der Prognosefehler um 50 % verringert wurde.

2 Um die Interpretation zu erleichtern, sollte die Zuordnung der Zahlenwerte zu den Merkmalsausprägungen so erfolgen, dass ein Anstieg der numerischen Werte auch mit einem Anstieg der inhaltlichen Ausprägung des Merkmals einhergeht.

Tabelle 7.9: Eckenkorrelation in einer 2x2-Tabelle

	Arbeiter	∼Arbeiter
SPD	100	50
Andere	0	50

N_D=0, N_C=100(50)=5000, $\gamma = 1$

Weil γ immer dann $\pm$ 1 wird, wenn es in der Tabelle keine diskordanten bzw. keine konkordanten Paare gibt, werden auch bei einer ‚Eckenkorrelation' perfekte Zusammenhänge ausgewiesen. In einer 2×2-Tabelle liegt eine Eckenkorrelation bereits bei einer unbesetzten Zelle vor (vgl. Tabelle 7.9). Beschränkt sich eine Hypothese auf das Wahlverhalten von Arbeitern (z. B. ‚Arbeiter wählen SPD'), dann ist das dargestellte Ergebnis, $\gamma = 1$, erwünscht. Gemäß dieser Hypothese liegt ein perfekter Zusammenhang vor. Anders stellt sich die Situation dar, wenn die Hypothese beinhaltet, dass Arbeiter überproportional häufig SPD wählen und Nicht-Arbeiter (∼Arbeiter) überproportional häufig andere Parteien. Ein perfekter Zusammenhang im Sinne dieser Hypothese wäre nur dann gegeben, wenn ausschließlich die Diagonale besetzt wäre. Kendalls τ_b (siehe unten) ist in diesem Fall ein angemesseneres Maß. τ_b beträgt für Tabelle 7.9 0,58.

Über die konkordanten und diskordanten Paare hinaus gibt es noch weitere Beziehungen zwischen Paaren in einer Kreuztabelle, die *ties* (Verknüpfungen). Insgesamt gibt es in jeder Tabelle $\frac{n(n-1)}{2}$ Paare, die sich aus der Zahl *konkordanter*, *diskordanter*, *in x verknüpfter*, *in y verknüpfter* und *in x und y verknüpfter* Paare zusammensetzen. Ein Paar ist in x verknüpft, wenn es in x dieselben Werte, in y aber unterschiedliche Werte hat. Eine Verknüpfung in y bedeutet denselben Wert in y, aber unterschiedliche Werte in x. In *x und y* ist ein Paar schließlich verknüpft, wenn dieselben Werte in x und y vorliegen, das Paar also in einer Zelle liegt.

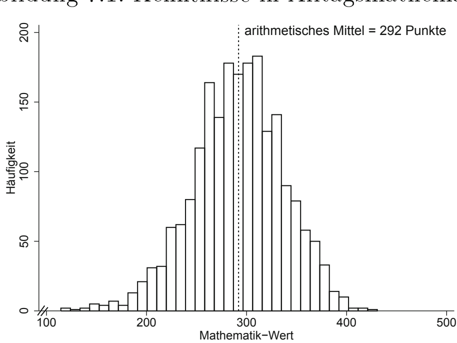
Kennwerte, die Verknüpfungen berücksichtigen, sind z. B. Kendalls tau-Maße (τ_a, τ_b, τ_c) und Somers' d. Kendalls τ-Maße und Somers' d unterscheiden sich im Zähler nicht von γ, dieser ist immer $N_C - N_D$. Im Nenner werden jedoch zusätzlich zu N_C und N_D Verknüpfungen berücksichtigt. Der Nenner ist bei diesen Maßen daher größer als bei γ. Aus diesem Grund nimmt γ systematisch höhere Werte an als die τ-Maße oder Somers' d. Nur wenn gar keine Verknüpfungen in der Tabelle vorliegen, sind die Werte identisch. Solange man berücksichtigt, dass γ systematisch

höhere Werte annimmt als die τ-Maße und Somers' d, spricht nichts gegen dessen Verwendung. Zudem ist der Betrag von γ als proportionale Fehlerreduktion interpretierbar.

7.5 Maß für ein nominalskaliertes und ein metrisches Merkmal: eta-Quadrat (η^2)

η^2 ist ein Maß für die Stärke des Zusammenhangs zwischen einem nominalskalierten unabhängigen Merkmal und einem mindestens intervallskalierten abhängigen Merkmal. Wir werden die Stärke des Zusammenhangs zwischen dem Geschlecht (nominal) und den Mathematikkenntnissen (metrisch) in der bundesdeutschen Bevölkerung untersuchen. Die Alltagsthese dazu lautet, dass es Unterschiede zwischen den Geschlechtern gibt, und Männern der Umgang mit Zahlen leichter fällt als Frauen. Zur Überprüfung der These verwenden wir Daten eines Leistungstests der bundesdeutschen erwachsenen Bevölkerung, die 1994 im Rahmen einer international angelegten Untersuchung erhoben wurden – den International Adult Literacy and Lifeskills Survey (IALS). Vom Design ähnelt IALS den PISA-Untersuchungen. Getestet wurde aber kein schulisches Wissen, sondern Grundkompetenzen in verschiedenen Bereichen, unter anderem im Umgang mit Zahlen (*numeracy*, Alltagsmathematik). In Abbildung 8.2 sind die Mathematikkenntnisse als Histogramm dargestellt. Die durchschnittlichen Kenntnisse (arithmetisches Mittel) lagen 1994 bei 292 Punkten auf einer Skala von 0 bis 500.

Abbildung 7.1: Kenntnisse in Alltagsmathematik



IALS. Deutschland 1994. n=2.062

η^2 ist ebenso wie λ (vgl. Kapitel 7.3.2) ein PRE-Maß. Zunächst wird versucht, die Ausprägung des abhängigen Merkmals ohne Kenntnis eines weiteren Merkmals vorherzusagen (Schritt 1). Für ein metrisches Merkmal ist der beste Prognosewert dessen *arithmetisches Mittel*. Die Summe der quadrierten Abweichungen aller Messwerte vom arithmetischen Mittel ist minimal, wie wir aus Kapitel 6.1 wissen. Bei Verwendung des arithmetischen Mittels als Schätzwert begehen wir deshalb den kleinsten (quadratischen) Fehler. Die Größe des Fehlers entspricht der Summe der Abweichungsquadrate (abgekürzt: SAQ, vgl. Kapitel 6.2). Diese lässt sich aus den Einzelmesswerten oder aus der Varianz s^2 berechnen.

- Für die 2062 Beobachtungen kann die Summe der Abweichungsquadrate mit einem Statistik-Programm aus den Einzelmesswerten berechnet werden:

$$\text{SAQ}_{ges} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{2062}(x_i - 292)^2 = 4222956\,.$$

- Alternativ kann die Summe der Abweichungsquadrate leicht aus der Varianz s^2 und der Zahl der Beobachtungen n rückgerechnet werden. Die Varianz $s^2 = \frac{\sum(x_i-\bar{x})^2}{n} = \frac{SAQ}{n}$ der Kenntnisse in Alltagsmathematik beträgt 2048 Punkte. Die Summe der Abweichungsquadrate beläuft sich daher auf

$$\text{SAQ}_{ges} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2 \cdot n = 2048 \times 2062 = 4222955\,.$$

  Mit Ausnahme von Rundungsfehlern sind die Werte identisch.

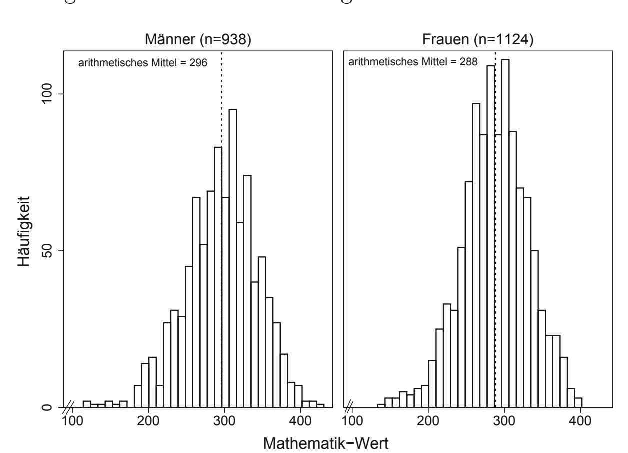
Im Beispiel beträgt die SAQ 4.222.956. Wir bezeichnen diese als Gesamtsumme der Abweichungsquadrate SAQ_{ges}, da alle Merkmalsträger in die Berechnung einfließen (Fehler_1).

Zur Prognose der abhängigen Variablen soll nun eine unabhängige Variable herangezogen werden (Schritt 2). Das Ausmaß der Verkleinerung des Vorhersagefehlers durch die unabhängige Variable gibt an, wie stark der Zusammenhang zwischen den beiden Variablen ist. Für jede Ausprägung der unabhängigen Variablen (Männer und Frauen) wird der Wert

der abhängigen Variablen (Mathematikkenntnisse) nun getrennt prognostiziert.

Die Mathematikkenntnisse sind in Abbildung 7.2 getrennt für Männer und Frauen dargestellt. Die durchschnittlichen Mathematikkenntnisse betragen 296,14 Punkte für die $n = 938$ befragten Männer, bei einer Varianz s^2 von 2242,66. Für die $n = 1124$ befragten Frauen belaufen sich die Mathematikkenntnisse auf durchschnittlich 288,14 Punkte bei einer Varianz s^2 von 1856,46. Die Frauen schneiden also geringfügig (um 8 Punkte) schlechter ab als die Männer.

Abbildung 7.2: Kenntnisse in Alltagsmathematik nach Geschlecht



IALS. Deutschland 1994.

Wir prognostizieren nun für Männer 296,14 Punkte und für Frauen 288,14 Punkte auf der Skala. Wie man in Abbildung 7.2 sieht, weichen die Mathematikkenntnisse der Frauen und Männer von ihrem jeweiligen Gruppenmittelwert ab. Die Summe der Abweichungsquadrate wird jetzt getrennt für Männer und Frauen berechnet:

$$\text{SAQ}_{Männer} = \sum_{i=1}^{938}(x_i - 296,14)^2 = 2103617 \quad \text{und}$$

$$\text{SAQ}_{Frauen} = \sum_{i=1}^{1124}(x_i - 288,14)^2 = 2086659.$$

Im Beispiel beträgt die SAQ bei den Männern 2.103.617 und bei den Frauen 2.086.659. Die Summe dieser beiden Werte entspricht dem Fehler, der bei der Vorhersage der Mathematikkenntnisse bei Kenntnis des Geschlechts begangen wird (Fehler$_2$): $\text{SAQ}_{kat} = 2103617 + 2086659 =$ 4.190.276.

Der letzte Schritt besteht in der Ermittlung des PRE-Maßes (Schritt 3). Die Maßzahl η^2 berechnet sich nun einfach aus der Differenz von SAQ_{ges} und SAQ_{kat} dividiert durch SAQ_{ges}:

$$\eta^2 = \frac{\text{Fehler}_1 - \text{Fehler}_2}{\text{Fehler}_1} = \frac{\text{SAQ}_{ges} - \text{SAQ}_{kat}}{\text{SAQ}_{ges}}. \tag{7.14}$$

Im Beispiel ergibt sich:

$$\eta^2 = \frac{\text{SAQ}_{ges} - \text{SAQ}_{kat}}{\text{SAQ}_{ges}} = \frac{4222956 - 4190276}{4222956} = 0{,}008\,.$$

η^2 hat einen Wertebereich von 0 (kein Zusammenhang) bis +1 (perfekter Zusammenhang). Ein Zusammenhang der Stärke 0,008 ist zu vernachlässigen (kein Zusammenhang). Da das unabhängige Merkmal nominalskaliert ist, ist η^2 vorzeichenlos. η^2 kann – mit 100 multipliziert – prozentual interpretiert werden. Der Fehler bei Vorhersage der Mathematikkenntnisse wird durch die Kenntnis des Geschlechts der Befragten um $0{,}008 \cdot 100 = 0,8\%$ verkleinert.

Gelegentlich wird auch die Quadratwurzel aus η^2 als Maß der Stärke des Zusammenhangs angegeben:

$$\eta = \sqrt{\eta^2}\,. \tag{7.15}$$

η kann ebenfalls Werte zwischen 0 und +1 annehmen. Im Beispiel resultiert

$$\eta = \sqrt{\eta^2} = \sqrt{0{,}0077} = 0{,}09\,.$$

Auch η ist nahe null. Bei den Befragten des IALS 1994 ist das Geschlecht zur Erklärung unterschiedlich guter Mathematikkenntnisse bedeutungslos.

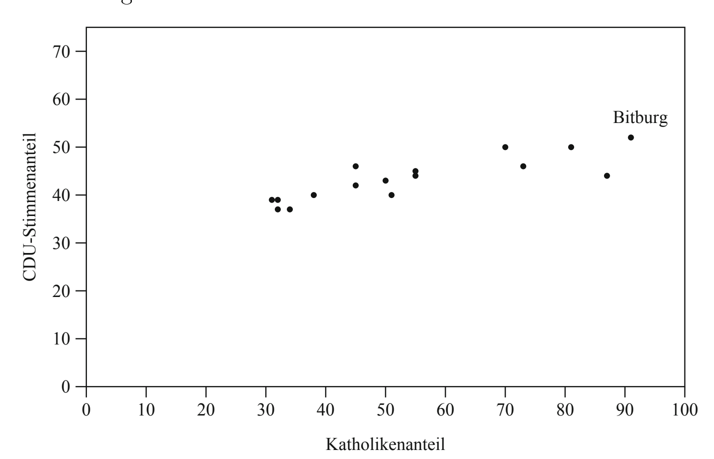
7.6 Maße für zwei metrische Merkmale: Kovarianz und Produkt-Moment-Korrelation

Kovarianz und Produkt-Moment-Korrelation messen die Stärke des *linearen* Zusammenhangs zwischen zwei mindestens intervallskalierten (metrischen) Merkmalen. Die Produkt-Moment-Korrelation wird nach dem Statistiker Karl Pearson auch als Pearsons r bezeichnet. Wenn ohne nähere Angabe von Korrelation gesprochen wird, dann ist meistens der Pearson'sche Korrelationskoeffizient gemeint.

Auch hier soll die Berechnung wieder an einem Beispiel verdeutlicht werden. Die CDU wurde nach dem Zweiten Weltkrieg als überkonfessionelle Partei gegründet. Da sie das „Erbe" der katholischen Zentrumspartei antrat, liegt die Vermutung nahe, dass auch die CDU besonders in katholischen Gebieten verankert ist, was sich in den Wahlergebnissen niederschlagen müsste. Unsere Hypothese lautet: „Je höher der Anteil der Katholiken in einem Bundestagswahlkreis, umso höher ist der Stimmenanteil der CDU". Die Hypothese soll anhand der amtlichen Ergebnisse der Bundestagswahl 1994 für die 16 rheinland-pfälzischen Bundestagswahlkreise überprüft werden. Merkmalsträger sind hier also nicht Personen, sondern Wahlkreise. Für jeden der 16 Wahlkreise liegt ein Messwertpaar vor, dass aus dem Katholikenanteil (x_i) und dem Stimmenanteil der CDU (y_i) besteht. Da es sich hier um Prozentwerte handelt, sind beide Merkmale ratioskaliert.

Der Zusammenhang zwischen zwei metrischen Merkmalen lässt sich in einem Streudiagramm darstellen. In Abbildung 7.3 ist der Zusammenhang zwischen dem Anteil der Katholiken und dem (Zweit-)Stimmenanteil der CDU dargestellt. Auf der x-Achse ist der Katholikenanteil eines Wahlkreises, auf der y-Achse der Stimmenanteil der CDU (an gültigen Stimmen) abgetragen. Beispielsweise betrug der Katholikenanteil im Wahlkreis 151 (Bitburg) 91,4 %, und die CDU erhielt dort knapp 53 % der gültigen Zweitstimmen.

Abbildung 7.3: Stimmenanteil der CDU und Katholikenanteil



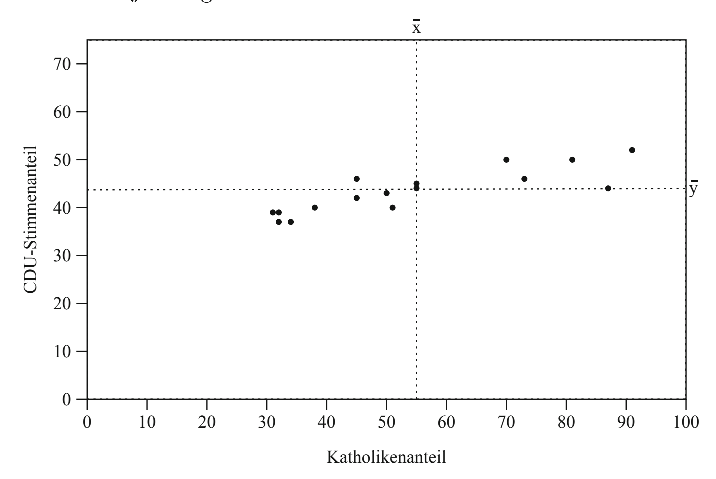
Auch ohne ein Zusammenhangsmaß zu berechnen, sieht man bereits, dass der Stimmenanteil der CDU umso höher ausfällt, je größer der Katholikenanteil ist. Zwischen den beiden Merkmalen besteht also ein positiver Zusammenhang, d. h. wenn x_i einen kleinen Wert annimmt, nimmt auch y_i einen kleinen Wert an, wenn x_i groß ist, ist auch y_i groß. Ein negativer Zusammenhang besteht dann, wenn die y_i-Werte mit zunehmenden x_i-Werten kleiner werden. Beispiel: Je höher der Arbeiteranteil in einem Wahlkreis, umso *schlechter* das Wahlergebnis der CDU. Kein Zusammenhang besteht, wenn eine Veränderung des x_i-Wertes die Größe des y_i-Wertes nicht beeinflusst.

Ob ein x_i- bzw. y_i-Wert groß oder klein ist, kann nicht absolut, sondern nur relativ zu allen anderen x_i- bzw. y_i-Messwerten bestimmt werden: Ein

Stimmenanteil von 38 % für die CDU wäre ein kleiner Wert verglichen mit den Wahlergebnissen der CDU in allen anderen 15 Wahlkreisen. Ebenso ist ein Katholikenanteil von 35 % in den 16 rheinland-pfälzischen Wahlkreisen nicht sehr hoch, während der gleiche Prozentsatz in Schleswig-Holstein ein hoher Wert wäre. Große Messwerte sind daher Messwerte, die überdurchschnittlich sind, kleine Messwerte solche, die unterdurchschnittlich ausfallen.

Die CDU erzielte im Durchschnitt in den 16 Wahlkreisen 43,96 % der gültigen Stimmen, und der durchschnittliche Katholikenanteil betrug 54,99 %. Große CDU-Werte sind also größer als 43,96 %, große Katholikenanteile sind größer als 54,99 % Katholiken. Wenn ein positiver Zusammenhang besteht, dann müßte ein überdurchschnittlicher Katholikenanteil auch ein überdurchschnittliches Stimmergebnis der CDU nach sich ziehen, ein unterdurchschnittlicher Katholikenanteil dementsprechend ein unterdurchschnittliches Wahlergebnis der CDU. Zeichnet man die arithmetischen Mittel $\bar{x}$ und $\bar{y}$ in die Graphik ein, erhält man vier Quadranten (vgl. Abbildung 7.4).

Abbildung 7.4: Stimmenanteil der CDU und Katholikenanteil mit den jeweiligen Mittelwerten

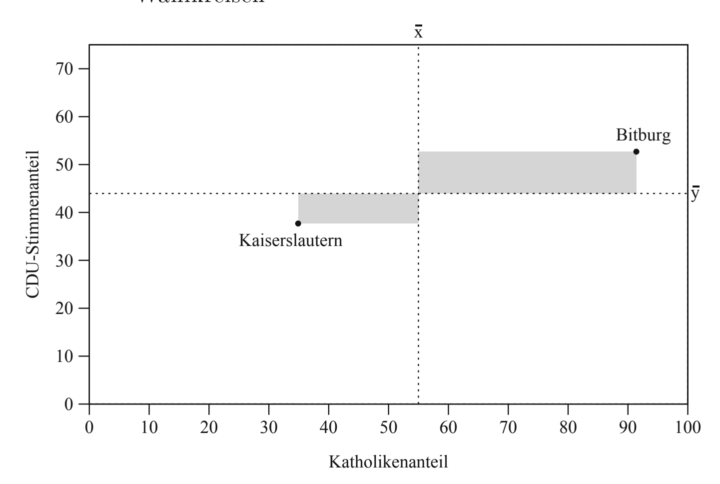


Im linken unteren Quadranten liegen alle Wahlkreise, die einen unterdurchschnittlichen Katholikenanteil und einen unterdurchschnittlichen

CDU-Stimmenanteil aufweisen. Im rechten oberen Quadranten befinden sich diejenigen Wahlkreise, die sowohl hinsichtlich des Katholikenanteils als auch des Stimmenanteils der CDU überdurchschnittlich abschneiden. Liegen die Messwertpaare hauptsächlich in diesen beiden Quadranten, dann variieren der Katholikenanteil und das Stimmergebnis der CDU positiv miteinander. Mit Ausnahme eines Wahlkreises liegen alle Messwertpaare im linken unteren und rechten oberen Quadranten, was auf einen starken positiven Zusammenhang deutet. Bei einer negativen Korrelation müssten die Messwertpaare vor allem im linken oberen und rechten unteren Quadranten liegen, da die y_i-Werte dann mit größer werdenden x_i-Werten abnehmen müssten. Liegt keine Korrelation vor, dann sind die Messwertpaare relativ gleichmäßig über alle Quadranten verteilt.

Um die Stärke des Zusammenhangs zu berechnen, muss berücksichtigt werden, wie weit die einzelnen Messwertpaare mit den Koordinaten $(x_i; y_i)$ vom Schwerpunkt der Verteilung $(\bar{x}; \bar{y})$ abweichen. Dies tun wir, indem wir für jedes Messwertpaar zunächst die Differenzen $(x_i - \bar{x})$ und $(y_i - \bar{y})$ berechnen. Diese Abweichungen sind für die beiden Wahlkreise Bitburg und Kaiserslautern in Abbildung 7.5 dargestellt.

Abbildung 7.5: Stimmenanteil der CDU und Katholikenanteil in zwei Wahlkreisen



Im Wahlkreis Bitburg sind 91,4 % der Bevölkerung katholisch, die Abweichung $x_i - \bar{x}$ beträgt $91{,}4 - 54{,}99 = 36{,}41$ Prozentpunkte. Die CDU erhielt dort 52,68 % der gültigen Stimmen, also 8,72 Prozentpunkte mehr als im Durchschnitt aller Wahlkreise ($y_i - \bar{y} = 52{,}68 - 43{,}96$). Das Abweichungsprodukt entspricht $(x_i - \bar{x}) \cdot (y_i - \bar{y}) = 36{,}41 \cdot 8{,}72 = 317{,}5$. Graphisch kann das Abweichungsprodukt – wie in Abbildung 7.5 auf der vorherigen Seite – als Fläche dargestellt werden.[3] Die Differenz $(x_i - \bar{x})$ entspricht der waagerechten Ausdehnung, die Differenz $(y_i - \bar{y})$ der senkrechten Ausdehnung des Rechtecks. Im Wahlkreis Kaiserslautern ist der Katholikenanteil und der Stimmenanteil der CDU unterdurchschnittlich. Das Abweichungsprodukt beträgt hier $(x_i - \bar{x}) \cdot (y_i - \bar{y}) = (34{,}89 - 54{,}99) \cdot (37{,}68 - 43{,}96) = (-20{,}1) \cdot (-6{,}28) = 126{,}2$. Das Abweichungsprodukt ist also kleiner als im Wahlkreis Bitburg, was man bereits optisch an der Größe der Flächen erkennt.

Die Kovarianz ist der Durchschnitt der Summe der Abweichungsprodukte für alle Messwertpaare. Während die Varianz die Streuung eines Merkmals bezeichnet, gibt die Kovarianz die *gemeinsame Streuung zweier Merkmale* an.

$$cov = \frac{\text{SAP}}{\text{Anzahl der Messwerte}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n} \qquad (7.16)$$

Zur Berechnung der Kovarianz verwenden wir eine Arbeitstabelle (Tabelle 7.10). In der ersten Spalte sind die Merkmalsträger – hier die Wahlkreise – verzeichnet, in der zweiten Spalte die Katholikenanteile (x_i) und in der dritten Spalte die Wahlergebnisse der CDU (y_i). Aus der zweiten und dritten Spalte lassen sich die arithmetischen Mittel des Katholikenanteils $\bar{x}$ und des Stimmenanteils für die CDU $\bar{y}$ ermitteln. Wenn dies geschehen ist, können die Abweichungen $(x_i - \bar{x})$ und $(y_i - \bar{y})$ und daraus das Abweichungsprodukt $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ für jeden einzelnen Wahlkreis berechnet werden. Lediglich in einem Wahlkreis – Neustadt-Speyer – ist das Abweichungsprodukt negativ (unterdurchschnittlicher Katholikenanteil bei überdurchschnittlichem CDU-Anteil). Die Summe der Abweichungsprodukte SAP beträgt im Beispiel 1189,13 ($n = 16$) und die Kovarianz

3 Für Messwertpaare im linken oberen und rechten unteren Quadranten gibt die Größe der Fläche den Betrag des Abweichungsprodukts an.

$$cov = \frac{1189,13}{16} = 74,32\,. \qquad (7.17)$$

Die Kovarianz ist null, wenn kein Zusammenhang besteht. Ein wesentlicher Nachteil der Kovarianz besteht darin, dass ihre Größe vom gewählten Maßstab abhängig ist: Hätten wir die beiden Merkmale nicht in Prozent, sondern in relativen Häufigkeiten gemessen, dann würde die Summe der Abweichungsprodukte um den Faktor 10.000 kleiner ausfallen. Diese Maßstabsabhängigkeit erschwert den Vergleich verschiedener Kovarianzen. Der Betrag der Kovarianz kann maximal so groß wie das Produkt der Standardabweichungen werden: $|cov| \leq s_x \cdot s_y$.

Durch Standardisierung entgeht man diesem Problem. Die Standardisierung erfolgt, indem man die Kovarianz durch ihr Maximum $cov_{max} = s_x \cdot s_y$ dividiert. Auf diese Weise erhält man die Produkt-Moment-Korrelation, die auch als Pearsons r bezeichnet wird:

$$r = \frac{cov_{xy}}{s_x \cdot s_y} = \frac{\frac{\text{SAP}}{n}}{\sqrt{\frac{\text{SAQ}_x}{n}} \cdot \sqrt{\frac{\text{SAQ}_y}{n}}} \quad . \qquad (7.18)$$

Aus der rechten Gleichung kann man n herauskürzen, so dass man auch schreiben kann:

$$r = \frac{\text{SAP}}{\sqrt{\text{SAQ}_x \cdot \text{SAQ}_y}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}\,. \qquad (7.19)$$

Der Wertebereich von r liegt zwischen -1 und $+1$. Bei den Extremwerten ± 1 liegen alle Messwerte auf einer Geraden. Nimmt r den Wert 0 an, dann besteht kein linearer Zusammenhang zwischen den beiden Merkmalen. r ist ebenso wie die Kovarianz ein symmetrisches Maß.

Zur Berechnung von r wird auf die Arbeitstabelle zurückgegriffen (Tabelle 7.10, S. 171). Die Summe der Abweichungsprodukte beträgt 1189,13.

Tabelle 7.10: Arbeitstabelle zur Berechnung von Kovarianz und r

Wahlkreis	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Neuwied	55,55	44,21	0,56	0,25	0,14	0,31	0,06
Ahrweiler	81,99	50,13	27,00	6,17	166,59	729,00	38,07
Koblenz	73,14	46,60	18,15	2,64	47,92	329,42	6,97
Cochem	70,78	50,94	15,79	6,98	110,21	249,32	48,72
Bad Kreuznach	32,60	39,10	−22,39	−4,86	108,82	501,31	23,62
Bitburg	91,40	52,68	36,41	8,72	317,50	1325,69	76,04
Trier	87,97	44,82	32,98	0,86	28,36	1087,68	0,74
Montabaur	50,76	43,42	−4,23	−0,54	2,28	17,89	0,29
Mainz	51,36	40,86	−3,63	−3,10	11,25	13,17	9,61
Worms	32,81	37,99	−22,18	−5,97	132,41	491,95	35,64
Frankenthal	31,98	39,71	−23,01	−4,25	97,79	529,46	18,06
Ludwigshafen	38,01	40,86	−16,98	−3,10	52,64	288,32	9,61
Neustadt-Speyer	45,61	46,48	−9,38	2,52	−23,64	87,98	6,35
Kaiserslautern	34,89	37,68	−20,10	−6,28	126,23	404,01	39,44
Pirmasens	45,98	42,79	−9,01	−1,17	10,54	81,18	1,37
Südpfalz	55,07	45,09	0,08	1,13	0,09	0,01	1,28
	$\bar{\mathbf{x}}$ = **54,99**	$\bar{\mathbf{y}}$ = **43,96**			**SAP = 1189,13**	$\mathbf{SAQ_x}$ = **6136,70**	$\mathbf{SAQ_y}$ = **315,87**

Die Abweichungsquadrate für den Katholikenanteil (SAQ_x) und die Abweichungsquadrate Stimmenanteil der CDU (SAQ_y) werden in den beiden letzten Spalten berechnet. Durch Einsetzen in Gleichung 7.19 erhält man:

$$r = \frac{\text{SAP}}{\sqrt{\text{SAQ}_x \cdot \text{SAQ}_y}} = \frac{1189{,}07}{\sqrt{6136{,}70 \cdot 315{,}96}} = 0{,}85\,.$$

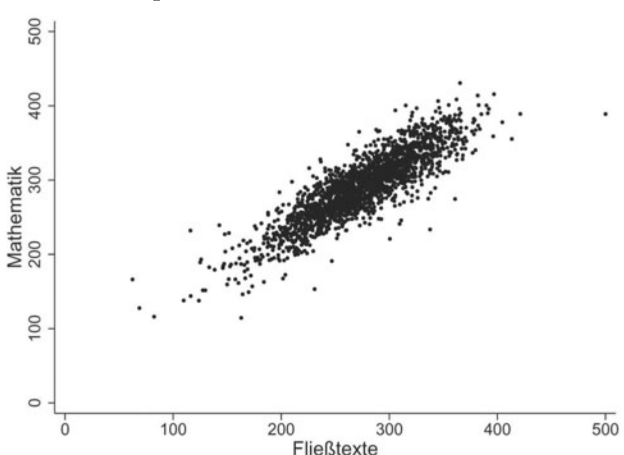
Liegen Kovarianz und Standardabweichungen der beiden Merkmale vor, dann kann r alternativ mit Gleichung 7.18 ermittelt werden. Die Kovarianz zwischen dem Katholikenanteil und dem Stimmenanteil der CDU beträgt 74,32, die Standardabweichung des Katholikenanteils $s_x = \sqrt{SAQ_x/n} = \sqrt{6136{,}7/16} = 19{,}58$ und die Standardabweichung des Stimmenanteils der CDU $s_x = \sqrt{SAQ_y/n} = \sqrt{315{,}87/16} = 4{,}44$.

$$r = \frac{cov_{xy}}{s_x \cdot s_y} = \frac{74,32}{19{,}58 \cdot 4{,}44} = 0{,}85$$

Das Ergebnis ist natürlich dasselbe. Die Korrelation zwischen dem Anteil der katholischen Bevölkerung und dem Wahlergebnis der CDU bei der Bundestagswahl 1994 in Rheinland-Pfalz beträgt also 0,85. Da dieser Wert sehr nah am Maximum 1 ist, liegt ein sehr starker Zusammenhang vor. r^2 kann – multipliziert mit 100 – prozentual interpretiert werden. $r^2 = 0,85^2 = 0,72$ heißt, dass 72 % der Unterschiede im Stimmenanteil der CDU auf den Katholikenanteil zurückgeführt werden können. Wir werden im nächsten Kapitel auf diese Interpretation zurückkommen.

Bei der Interpretation der Stärke des Zusammenhangs zwischen Katholikenanteil und Stimmenergebnis für die CDU muss berücksichtigt werden, dass mit Aggregatdaten häufig stärkere Zusammenhänge gemessen werden als mit Individualdaten, was auf Gruppierungseffekte zurückgeführt werden kann (vgl. Pappi 1977, 90). Bei Individualdaten beobachtet man nur selten so starke Zusammenhänge. Ein Beispiel ist der Zusammenhang zwischen den Lesekenntnissen (Fließtexte) und den Mathematikkenntnissen, der in Abbildung 7.6 dargestellt ist. Hier ist $r = 0,88$. Der enge Zusammenhang widerspricht der von vielen Menschen präferierten These

Abbildung 7.6: Lese- und Mathematikkenntnisse

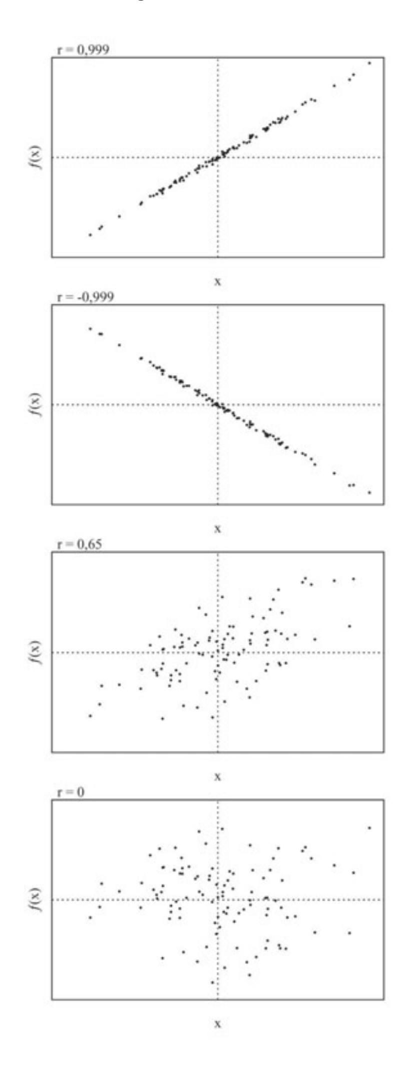


IALS 1994. Deutschland. n=2062.

wonach Fähigkeiten im Umgang mit Texten nur schwach mit Fähigkeiten im Umgang mit Zahlen korrelieren.

In Abbildung 7.7 auf der nächsten Seite sind unterschiedlich stark ausgeprägte Zusammenhänge dargestellt. In der obersten Abbildung ist ein positiver Zusammenhang der Stärke 0,999 abgebildet, direkt darunter ein negativer Zusammenhang gleicher Intensität. In der dritten Abbildung korrelieren die Merkmale immer noch recht stark (r=0,65), während in der untersten Abbildung kein Zusammenhang zwischen den beiden Merkmalen besteht (r=0). r erfasst ausschließlich den *linearen* Zusammenhang zwischen zwei Merkmalen. Auch bei einem nichtlinearen (z. B. u-förmigen) Zusammenhang zwischen zwei Merkmalen kann r einen schwachen bzw. keinen Zusammenhang ausweisen.

Abbildung 7.7: Darstellung unterschiedlich hoher Korrelationen



Aufgaben zu Zusammenhangsmaßen

1. Sie möchten den Zusammenhang zwischen der Konfession und der Wahlabsicht prüfen. In der nachstehenden Tabelle ist die Wahlabsicht für Nicht-Katholiken und Katholiken wiedergegeben.
   Bitte berechnen Sie die Spalten- und Zeilenprozente und interpretieren Sie die Aussage der Tabelle! Wie stark ist der Zusammenhang zwischen beiden Merkmalen ausgeprägt? Berechnen Sie bitte den Kontingenzkoeffizienten C, Cramérs V sowie λ (Vorhersage der Wahlabsicht)!

	nicht katholisch	katholisch	Summe
CDU/CSU	236	297	533
SPD	390	205	595
ANDERE	268	179	447
Summe	894	681	1575

Quelle: ALLBUS 1994, westdeutsche Befragte

2. Prüfen Sie anhand der abgebildeten Tabelle, ob es ein Zusammenhang zwischen dem Schulabschluss der Interviewer und dem Schulabschluss der Befragten besteht. Berechnen Sie bitte ein angemessenes Zusammenhangsmaß.

	Interviewer			
Befragter	Hauptschule	Realschule	FHR/ ABI	Summe
Hauptschule	389	591	670	1650
Realschule	162	352	503	1017
FHR/Abitur	107	227	317	651
Summe	658	1170	1490	3318

Quelle: ALLBUS 1994, westdeutsche Befragte

3. Aus Erfahrung wissen Sie, dass ältere Menschen nur ungern Männern die Tür aufmachen. Sie vermuten deshalb, dass in einer Umfrage die männlichen Interviewer eher junge Menschen befragt haben, die weiblichen Interviewer eher ältere Menschen. Nachfolgend ist das Durchschnittsalter aller Befragten wiedergegeben, das Durchschnittsalter der von den Interviewern Befragten und das Durchschnittsalter der von den Interviewerinnen Befragten. Berechnen Sie auch hier bitte ein angemessenes Zusammenhangsmaß.

	Altersdurchschnitt	Varianz	n
Alle Befragte	45,8356	286,1653	3442
Von Mann Interviewte	45,6635	282,8915	2320
Von Frau Interviewte	46,1916	293,0079	1121

Quelle: ALLBUS 1994

4. Sie möchten den Zusammenhang zwischen der Außentemperatur und Ihrem Eiskonsum feststellen. Dazu haben Sie an fünf aufeinanderfolgenden Tagen die Temperatur sowie die Anzahl der von Ihnen verzehrten Eis notiert.
Bitte berechnen Sie die Stärke des Zusammenhangs mit Hilfe des Pearsonschen Korrelationskoeffizienten r und interpretieren Sie das Ergebnis! Zeichnen Sie die Messwertpaare in ein Diagramm ein!

Tag	Temperatur (°Celsius)	Eiskonsum (Anzahl)
1	15	1
2	30	7
3	20	2
4	24	4
5	17	2

5. Bitte antworten Sie mit richtig oder falsch.
   a) Das Tauschen von Spalten in einer Kreuztabelle hat einen Einfluss auf den Wert von γ.
   b) Das Tauschen von Spalten in einer Kreuztabelle hat einen Einfluss auf den Wert von Cramérs V.

8 Lineare Regression

8.1 Grundgedanke der Regressionsanalyse

Mit Hilfe einer Regressionsanalyse untersucht man den Einfluss von einer oder mehreren unabhängigen Variablen auf *eine einzige* abhängige Variable. Das Verfahren heißt *Regressionsanalyse*, weil die Ausprägung der abhängigen Variable auf die Ausprägungen der unabhängigen Variablen *zurückgeführt* („regrediert") wird. Wir beschränken uns für diese Einführung auf die *lineare Einfach*regression.

Bei einer linearen Regression wird eine Beziehung zwischen unabhängiger und abhängiger Variable unterstellt, die sich durch eine *Gerade* darstellen lässt. Eine solche lineare Beziehung könnte heißen: Y ist immer um drei Einheiten größer als X. Mathematisch formuliert: $y = x + 3$. Das Modell könnte aber auch heißen: Y ist immer um den Faktor 250 größer als X: $y = 250 \cdot x$. Ob die Beziehung zwischen zwei metrischen Merkmalen linear ist, lässt sich mit einem Streudiagramm leicht überprüfen. Voraussetzung einer linearen Regression ist, dass die beteiligten Merkmale *metrisches Skalenniveau* aufweisen.[1] Bei einer *Einfach*regression (bivariate Regression) wird der Einfluss *einer einzigen* unabhängigen Variable auf die abhängige Variable geschätzt. Bei einer multiplen Regression wird dagegen der Einfluss mehrerer Merkmale auf eine abhängige Variable betrachtet.

Die Formulierung „Regression von Y auf X" gibt die Erklärungsrichtung an. X wird zur Erklärung von Y herangezogen; X ist die unabhängige Variable, Y ist die abhängige Variable. In der Sprechweise der linearen Regression sagt man dann: „Man führt die Ausprägung des Merkmals Y auf die Ausprägung des Merkmals X zurück", deshalb Regression *von* Y

1 Im Modell können auch dichotome *un*abhängige Merkmale berücksichtigt werden, die abhängige Variable muss metrisch sein.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_8,

auf X. Welche Variable als abhängig und welche als unabhängig betrachtet wird, ist von der Fragestellung abhängig, wie bereits in Kapitel 5.1.2 erläutert wurde.

8.2 Das mathematische Modell der linearen Regression

Das mathematische Modell einer linearen Einfachregression beinhaltet ein unabhängiges Merkmal x, ein abhängiges Merkmal y, die Konstante a und die Steigung b.

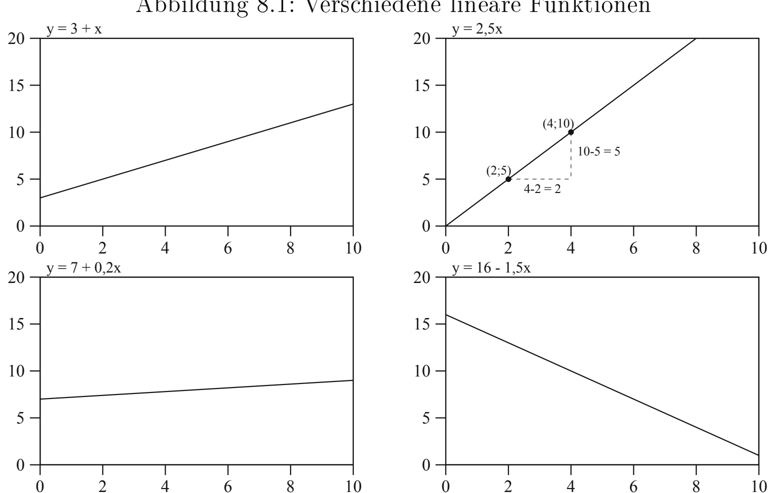
$$y = a + b \cdot x \tag{8.1}$$

Der Wert y bestimmt sich aus der Konstanten a zuzüglich des mit dem Faktor b multiplizierten Wertes x. Egal, welche Werte für a und b eingesetzt werden, das Ergebnis ist immer eine Gerade. Graphisch betrachtet ist a der Schnittpunkt der Geraden mit der y-Achse, b ist die Steigung der Geraden. Bei einer Steigung von $b = 0$ verläuft die Gerade parallel zur x-Achse ($y = a + 0 \cdot x = a$) oder es ist die x-Achse selbst (für $a = 0$). Abbildung 8.1 auf der nächsten Seite zeigt verschiedene Geraden für unterschiedliche Werte von a und b. Links oben in Abbildung 8.1 ist die Funktion $y = 3 + x$ wiedergegeben. Der Schnittpunkt mit der y-Achse ist also bei 3, die Steigung ist 1 ($1 \cdot x = x$). Rechts daneben ist die Funktion $y = 2{,}5 \cdot x$ dargestellt. Der Schnittpunkt mit der y-Achse ist 0, die Steigung 2,5. Eine solche Gerade wird auch „Ursprungsgerade“ oder „Nullpunktgerade“ genannt, da sie durch den Ursprung bzw. Nullpunkt des Koordinatensystems geht. Links unten ($y = 7 + 0{,}2 \cdot x$) ist eine Gerade mit der sehr geringen Steigung von 0,2 abgebildet. Die Gerade verläuft also fast parallel zur x-Achse. Daneben ($y = 16 - 1{,}5 \cdot x$) ist eine Gerade mit der Steigung $-1{,}5$ abgebildet, d. h. die Gerade steigt nicht, sondern sie fällt mit zunehmendem Wert von x.

Die Steigung der Geraden lässt sich auch immer als das Verhältnis einer Differenz zweier Punkte auf der y-Achse zur Differenz derselben Punkte auf der x-Achse angeben. Betrachten wir dazu die Gerade rechts oben in Abbildung 8.1. Zwischen zwei Punkten mit den Koordinaten $(x_1; y_1) = (2; 5)$ und $(x_2; y_2) = (4; 10)$ liegt die Differenz auf der y-Achse von $y_2 - y_1 = 10 - 5 = 5$ und auf der x-Achse von $x_2 - x_1 = 4 - 2 = 2$. Eine andere

Bezeichnung für denselben Sachverhalt ist $\Delta Y = 5$ und $\Delta X = 2$.[2] Dies wird auch als „Steigungsdreieck“ bezeichnet, da der Quotient $\Delta Y/\Delta X$ die Steigung der Geraden angibt, im Beispiel $5/2 = 2{,}5$. Dies bedeutet, dass Y um 2,5 Einheiten ansteigt, wenn X um eine Einheit steigt.

Abbildung 8.1: Verschiedene lineare Funktionen



Dieses *lineare Modell* wird uns immer wieder begegnen. Als statistisches Modell wenden wir es an, wenn wir einen linearen Zusammenhang zwischen zwei Variablen unterstellen und aufgrund dieses Zusammenhangs eine Prognose der abhängigen Variablen abgeben wollen.

8.3 Bestimmung der Regressionsfunktion

Zur Illustration greifen wir das Beispiel aus Kapitel 7.6 auf. Wir möchten bestimmen, wie stark der Katholikenanteil in einem Wahlkreis das Wahlergebnis der CDU beeinflusst, und ein Modell berechnen, das es uns erlaubt, den Stimmenanteil der CDU auf Grundlage des Katholikenanteils zu schätzen. Dazu führen wir eine lineare Regression des CDU-Stimmenanteils auf den Anteil der Katholiken in den rheinland-pfälzischen

2 Δ ist das griechische große Delta und wird häufig für die Bezeichnung eines Intervalls benutzt, in diesem Fall also für eine Strecke.

Bundestagswahlkreisen durch, d. h. wir suchen eine Gerade zur Vorhersage des CDU-Stimmenanteils.

Die allgemeine Funktion dieser Geraden geht aus Gleichung 8.1 hervor, wobei der Schnittpunkt der Geraden mit der y-Achse, a, bei der Regression als *Regressionskonstante*, und die Steigung der Regressionsgeraden, b, als *Regressionskoeffizient* oder Regressionsgewicht bezeichnet wird. Wenn Gleichung 8.1 solchermaßen als „Schätzmodell" verwendet wird, schreibt man sie als:

$$\hat{y}_i = a + b \cdot x_i \,. \tag{8.2}$$

Die Schreibweise $\hat{y}_i$ (sprich: y-Dach) verwendet man, um deutlich zu machen, dass es sich bei $\hat{y}_i$ um eine Schätzung aufgrund dieser Gleichung handelt und nicht um einen beobachteten Wert. Die beobachteten Werte y_i weichen ja mehr oder weniger von der Geraden $\hat{y}_i$ ab. Die Abweichungen e_i werden auch als Residuen bezeichnet.

$$e_i = y_i - \hat{y}_i \tag{8.3}$$

Die Ermittlung der Abstände erfolgt graphisch gesehen immer entlang der Richtung der abhängigen Variablen, da es darum geht, bei der Vorhersage dieser Variablen möglichst wenige Fehler zu machen. Bei einer Regression von Y auf X werden die Abstände daher entlang der Ausprägung der Y-Variablen minimiert. Würde man die Abstände entlang der X-Variablen bestimmen, käme dies einer Umkehrung der Richtung der Beziehung zwischen den Variablen gleich, so dass X nicht mehr die unabhängige, sondern die abhängige Variable wäre und Y die unabhängige anstatt der abhängigen Variablen. Die Geraden, die sich auf diesen beiden Wegen ermitteln lassen, sind nicht identisch; deshalb ist genau darauf zu achten, welche Variable die abhängige und welche die unabhängige ist (vgl. Clauß und Ebner 1989, S. 108–112).

Da die Vorhersage natürlich möglichst gut sein soll, stellt sich die Frage, welche Gerade die Punktewolke (vgl. Abbildung 7.3 auf Seite 166) am besten beschreibt. Man könnte eine Gerade „per Augenschein" durch die

Punkte legen. Diese würde die Lage der Punkte vermutlich nur sehr unzureichend wiedergeben. Naheliegend scheint es, die Gerade zu suchen, bei der die Summe der Abweichungen von der Geraden null ist ($\sum e_i = 0$). Dies ist allerdings kein geeignetes Kriterium, weil es in jeder Punktwolke mehrere Geraden gibt, die diese Bedingung erfüllen. Statt dessen minimiert man die Summe der quadrierten Abstände $\sum e_i^2$, weshalb das Verfahren auch als **Kleinste-Quadrate-Methode** bezeichnet wird (OLS = Ordinary Least Squares),

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a + bx_i)^2 = min! \tag{8.4}$$

wobei die $\hat{y}_i$-Werte die aufgrund der (noch zu bestimmenden) Geraden ermittelten Schätzwerte und y_i die beobachteten Werte sind.

Aus der linearen Algebra ist vielleicht noch bekannt, dass man das Minimum einer Funktion erhält, wenn man die 1. Ableitung null setzt und die 2. Ableitung bei einem Minimum positiv sein muss (ist die 2. Ableitung negativ, erhält man ein Maximum). Da die beiden Parameter a und b gesucht werden, muss Gleichung 8.4 partiell nach a und b abgeleitet werden (vgl. Bortz 2004, S. 185 f.). Für b erhält man nach einigen Umformungen folgende Formel:

$$b = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum\limits_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{SAP}}{\text{SAQ}_x}\,. \tag{8.5}$$

Zur Berechnung von b benötigen wir die Summe der Abweichungsprodukte (SAP) und die Summe der Abweichungsquadrate von x (SAQ_x).

Aus b und den arithmetischen Mitteln $\bar{x}$ und $\bar{y}$ kann dann die Regressionskonstante a bestimmt werden. Eine Eigenschaft der durch die Kleinste-Quadrate-Methode berechneten Geraden ist es nämlich, dass sie durch den Punkt ($\bar{x}$;$\bar{y}$) – den Schwerpunkt der Verteilung - verläuft. Außerdem ist bekannt, dass a der Schnittpunkt der Geraden mit der y-Achse ist, also verläuft die Gerade durch den Punkt (0;a). Damit haben wir zwei Punkte

der Geraden und können die Steigung der Geraden b auch als Steigungsdreieck festlegen:

$$b = \frac{\Delta y}{\Delta x} = \frac{\bar{y} - a}{\bar{x} - 0} = \frac{\bar{y} - a}{\bar{x}} \,. \tag{8.6}$$

Durch Umformen ergibt sich

$$a = \bar{y} - b \cdot \bar{x} \,. \tag{8.7}$$

Die Summe der Abweichungsprodukte SAP und die Summe der Abweichungsquadrate SAQ_x haben wir bereits in Kapitel 7.6 (Tabelle 7.10) berechnet. Durch Einsetzen in die Gleichungen 8.5 und 8.7 ergibt sich

$$b = \frac{\text{SAP}}{\text{SAQ}_x} = \frac{1189{,}13}{6136{,}70} = 0{,}194 \quad \text{und}$$

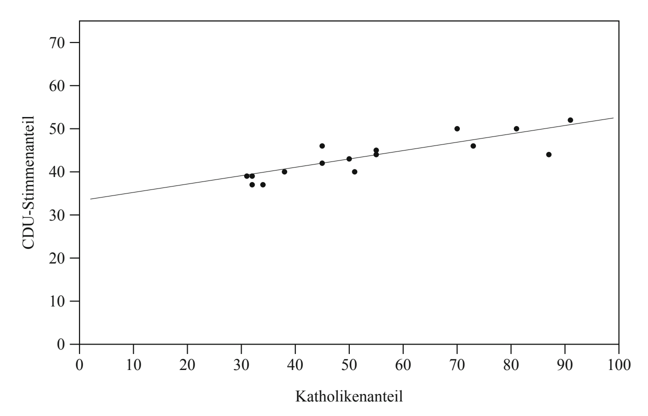
$$a = \bar{y} - b_{yx} \cdot \bar{x} = 43{,}96 - 0{,}194 \cdot 54{,}99 = 33{,}29 \,.$$

Die Regressionsgerade in unserem Beispiel lautet also:

$$\hat{y}_i = 33{,}29 + 0{,}194 \cdot x_i \,.$$

Regressionskonstante und Regressionskoeffizient lassen sich anschaulich interpretieren: Wenn die unabhängige Variable um eine Einheit ansteigt, ändert sich der Schätzwert der abhängigen Variable um b Einheiten. Für einen Anstieg des Katholikenanteils um einen Prozentpunkt prognostizieren wir einen Anstieg des CDU-Stimmenanteils um 0,194 Prozentpunkte. Entsprechend sagen wir einen Anstieg des CDU-Stimmenanteils von 1,94 Prozentpunkten bei einem Anstieg des Katholikenanteils um 10 Prozentpunkte vorher. Ist $b = 0$, dann übt die unabhängige Variable keinen Einfluss auf die abhängige Variable aus oder der Einfluss ist nichtlinear. Die


Abbildung 8.2: Regression des CDU-Stimmenanteils auf den Katholikenanteil



Regressionskonstante ist der Wert, den wir für $x_i = 0$ prognostizieren. Bei einem Katholikenanteil von 0 % prognostizieren wir einen Stimmenanteil von 33,29 % für die CDU. In Abbildung 8.2 ist die ermittelte Regressionsgerade eingezeichnet.

Die Gleichung $\hat{y}_i = 33{,}29 + 0{,}194 \cdot x_i$ kann nun zur Prognose („Schätzung“) der Y-Variablen aufgrund des Wertes der X-Variablen verwendet werden. Auf unser Beispiel angewendet, kann man jetzt also zu einem beliebigen Katholikenanteil in einem Wahlkreis x_i den Stimmenanteil der CDU $\hat{y}_i$ „schätzen“.

So würde man bei einem Katholikenanteil von 70,78 % aufgrund der Regressionsgleichung einen Stimmenanteil der CDU von $\hat{y} = 33{,}29 + 0{,}194 \cdot 70{,}78 = 47{,}02\,\%$ prognostizieren. Im Wahlkreis Cochem, wo genau dieser Katholikenanteil vorkommt, beträgt der *tatsächliche* Stimmenanteil der CDU aber 50,94 %, liegt also über dem geschätzten Wert. Differenzen zwischen $\hat{y}_i$- und y_i-Werten kommen vor, weil nicht alle beobachteten Werte exakt auf einer Geraden liegen.

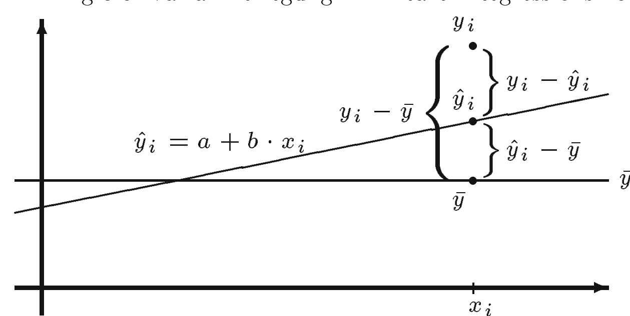
8.4 Qualität der Regression

Die Regressionsgerade repräsentiert die beobachteten Werte umso besser, je weniger die geschätzten Werte $\hat{y}_i$ von den beobachteten Werten y_i abweichen. Ein Maß für die Annäherung der Geraden an die beobachteten Werte und damit die *Prognosequalität* der Regressionsgleichung ist der *Determinationskoeffizient* R^2, ein dem Korrelationskoeffizienten r sehr eng verwandtes Maß. Andere Bezeichnungen für R^2 sind „Bestimmtheitsmaß" oder „Varianzaufklärung". Letztere Bezeichnung weist auf die R^2 innewohnende Logik hin. R^2 gibt an, *welcher Anteil der Varianz der abhängigen Variablen durch die unabhängige Variable erklärt wird.*

Zur Erklärung von R^2 sollte man sich noch einmal die Logik eines PRE-Maßes (vgl. Kapitel 7.3.2 und 7.5) vergegenwärtigen. Bei PRE-Maßen versucht man zunächst, die abhängige Variable ohne Hinzuziehung einer unabhängigen Variablen zu prognostizieren. Im Fall einer metrischen Variablen ist der beste Schätzwert deren arithmetisches Mittel $\bar{y}$ (=ursprüngliche Prognose), im Beispiel also das durchschnittliche Wahlergebnis der CDU in den rheinland-pfälzischen Bundestagswahlkreisen, nämlich 43,96 %. Die Differenzen $(y_i - \bar{y})$ geben an, wie weit wir mit dieser Prognose von den beobachteten Werten entfernt sind. Anschließend wird die abhängige Variable auf Basis einer unabhängigen Variable vorhergesagt. Der beste Schätzwert ist nun der durch die ermittelte Regressionsgleichung prognostizierte Wert $\hat{y}_i$ für eine konkrete Ausprägung der unabhängigen Variable x_i (=neue Prognose). Die Abweichungen $y_i - \hat{y}_i$ geben an, wie weit die Regressionsgerade von den beobachteten Werten entfernt ist. Das Gütekriterium der Prognose bestimmt sich nun daraus, in welchem Umfang die Fehler auf Basis der ursprünglichen Prognose durch die neue Prognose vermindert werden.

Diese Herleitung von R^2 wird anhand *einer* Beobachtung in Abbildung 8.3 auf der nächsten Seite grafisch veranschaulicht. x_i ist der Wert der unabhängigen Variable (Katholikenanteil). Zu ihm gehört der beobachtete Wert der abhängigen Variable y_i (Stimmenanteil der CDU). Der Wert auf der Parallelen zur x-Achse ist der Mittelwert der abhängigen Variablen $\bar{y}$ - im Beispiel der durchschnittliche Anteil der CDU. Also der Wert, den wir ohne Kenntnis einer unabhängigen Variablen vorhersagen. Der Wert auf der berechneten Geraden $\hat{y}_i = a + b \cdot x_i$ ist die Schätzung der abhängigen Variablen $\hat{y}_i$ durch die Regressionsgerade.

Abbildung 8.3: Varianzzerlegung im linearen Regressionsmodell



Die Abweichung des beobachteten Wertes vom Mittelwert $(y_i - \bar{y})$ soll erklärt werden. Sie lässt sich aufteilen in die Abweichung des beobachteten Wertes vom Schätzwert der Regressionsgeraden $(y_i - \hat{y}_i)$ *und* die Abweichung des Schätzwertes der Regressionsgeraden zum Mittelwert $(\hat{y}_i - \bar{y})$. Die Differenz $\hat{y}_i - \bar{y}$ kann mit Hilfe der Regression von Y auf X erklärt werden. Die Differenz $y_i - \hat{y}_i$ kann nicht auf X zurückgeführt werden, sie bleibt unerklärt.

$$\underbrace{y_i - \bar{y}}_{\text{zu erklärende Abweichung}} = \underbrace{y_i - \hat{y}_i}_{\text{nicht erklärte Abweichung}} + \underbrace{\hat{y}_i - \bar{y}}_{\text{erklärte Abweichung}} \tag{8.8}$$

Im Wahlkreis Cochem liegen wir mit der Prognose des durchschnittlichen CDU-Anteils $\bar{y}$ von 43,96 % um 6,98 Prozentpunkte daneben, denn der tatsächliche Stimmenanteil der CDU in Cochem y_i beträgt 50,94 %. Die zu erklärende Abweichung beträgt also $y_i - \bar{y} = 50{,}94 - 43{,}96 = 6{,}98$ Prozentpunkte. Zur Erklärung des Stimmenanteils der CDU ziehen wir den Katholikenanteil im Wahlkreis heran. Auf Basis der berechneten Regressionsgeraden $(\hat{y}_i = 33{,}29 + 0{,}194 \cdot x_i)$ erwarten wir für einen Wahlkreis mit einem Katholikenanteil von 70,78 % (Wert für Cochem), dass 47,02 % der Wähler für die CDU stimmen. Mit Hilfe des Katholikenanteils wird die Schätzung also besser, sie liegt näher am tatsächlichen Stimmenergebnis der CDU. Durch die Regression werden $\hat{y}_i - \bar{y} = 47{,}02 - 43{,}96 = 3{,}06$ Prozentpunkte des überdurchschnittlichen Stimmenanteils der CDU. Der

Katholikenanteil erklärt den Wahlerfolg der CDU in Cochem nicht vollständig, aber einen Teil davon. Die weiterhin nicht erklärte Abweichung beträgt $y_i - \hat{y}_i = 50{,}94 - 47{,}02 = 3{,}92$ Prozentpunkte, denn auch bei Kenntnis des Katholikenanteils prognostizieren wir 3,92 Prozentpunkte zu wenig.

Diese Abweichungen müssen nun für alle 16 Wahlkreise berechnet werden. Bevor sie summiert werden, müssen sie noch *quadriert* werden, denn sonst heben sich positive und negative Abweichungen auf, so dass die gesamte Abweichung für alle Wahlkreise 0 betragen würde.

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Gesamt-SAQ}_y} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Unerklärte-SAQ}_y} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Erklärte-SAQ}_y} \quad (8.9)$$

Die zu erklärende Gesamt-SAQ_y setzt sich also aus einer durch die Regression unerklärten SAQ_y und einer durch die Regression erklärten SAQ_y zusammen. Das Verhältnis der erklärten SAQ_y zur Gesamt-SAQ_y ist das Maß für die Güte der Regression, R^2.

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{Erklärte-SAQ}_y}{\text{Gesamt-SAQ}_y} \quad (8.10)$$

Zur Verdeutlichung der PRE-Maß-Logik kann man auch schreiben:

$$\begin{aligned} R^2 &= \frac{\text{Fehler}_1 - \text{Fehler}_2}{\text{Fehler}_1} \\ &= \frac{\text{Gesamt-SAQ}_y - \text{Unerklärte-SAQ}_y}{\text{Gesamt-SAQ}_y} \\ &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} . \end{aligned} \quad (8.11)$$

Die Gesamt-SAQ$_y$ in Gleichung 8.11 entspricht den Fehlern bei der ursprünglichen Prognose (Fehler$_1$), die Unerklärte-SAQ$_y$ den Fehlern auf Basis der neuen Prognose (Fehler$_2$). Die Differenz gibt die Verringerung der Fehler durch die Hinzuziehung der unabhängigen Variable an. Die Gleichungen 8.10 und 8.11 sind natürlich identisch, da Gesamt-SAQ$_y$ − Unerklärte-SAQ$_y$ = Erklärte-SAQ$_y$.

R^2 hat einen Wertebereich von 0 bis 1. R^2 nimmt den Wert 0 an, wenn die unabhängige Variable X die Vorhersage nicht verbessert. In diesem Fall ist auch $b = 0$. Je größer R^2 ist, desto größer ist der Anteil der erklärten Variation der abhängigen Variablen. Bei $R^2 = 1$ liegen alle Messwerte auf der Regressionsgeraden. Die Unterschiede in der abhängigen Variable lassen sich dann vollständig (zu 100 %) auf die unabhängige Variable zurückführen.

Um die Berechnung von R^2 in unserem Beispiel durchzuführen, erweitert man Tabelle 7.10 (S. 171) um die Spalten zur Berechnung von $\hat{y}_i$, $y_i - \hat{y}_i$, $(y_i - \hat{y}_i)^2$, $\hat{y}_i - \bar{y}$ und $(\hat{y}_i - \bar{y})^2$. Die entsprechende Tabelle 8.1 ist auf Seite 189 dargestellt.

Durch Einsetzen der x_i-Werte in die Regressionsgleichung lassen sich die $\hat{y}_i$-Werte berechnen. Die Summe der quadrierten Abweichungen $\sum(y_i - \hat{y}_i)^2$ ist die unerklärte SAQ von Y, die Summe der quadrierten Abweichungen $\sum(\hat{y}_i - \bar{y})^2$ ist die erklärte SAQ von Y. Die unerklärte SAQ beträgt 85,49, die erklärte SAQ 231,05 und die gesamte SAQ 315,87.[3] R^2 lässt sich nach Gleichung 8.10 wie folgt berechnen:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{erklärte SAQ}_y}{\text{Gesamt-SAQ}_y} = \frac{231{,}05}{315{,}87} = 0{,}73\,;$$

bzw. alternativ nach Gleichung 8.11:

3 Aufgrund von Rundungsungenauigkeiten entspricht die Summe aus erklärter und unerklärter SAQ_y $(85{,}49 + 231{,}05 = 316{,}53)$ nicht exakt der Gesamt-SAQ_y $(315{,}87)$. SPSS ermittelt eine unerklärte SAQ_y von $85{,}528$, eine erklärte SAQ_y von $230{,}341$ und eine Gesamt-SAQ_y von $315{,}869$.

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{315,87 - 85,49}{315,87} = 0,73\,.$$

$R^2 = 0{,}73$ ist ein sehr hoher Wert, der in der Praxis nur selten erreicht wird. Mit 100 multipliziert, lässt er sich als Prozentwert interpretieren: Die Varianzaufklärung beträgt 73 %. Oder anders ausgedrückt: 73 % der Unterschiede im Stimmenanteil der CDU lassen sich auf die Höhe des Katholikenanteils im Wahlkreis zurückführen. Und als PRE-Maß: Der Fehler bei Prognose des Stimmenanteils der CDU wird durch Kenntnis des Katholikenanteils um 73 % reduziert.

Aus R^2 lässt sich im bivariaten Fall der Korrelationskoeffizient r bestimmen, denn

$$r = \sqrt{R^2}\,. \tag{8.12}$$

Im Beispiel würde der auf diese Weise ermittelte Wert $r = 0{,}85$ betragen. R^2 ist allerdings nur bei einer Regression mit **einer** unabhängigen Variablen identisch mit dem quadrierten Korrelationskoeffizienten r^2 aus Kapitel 7.6.

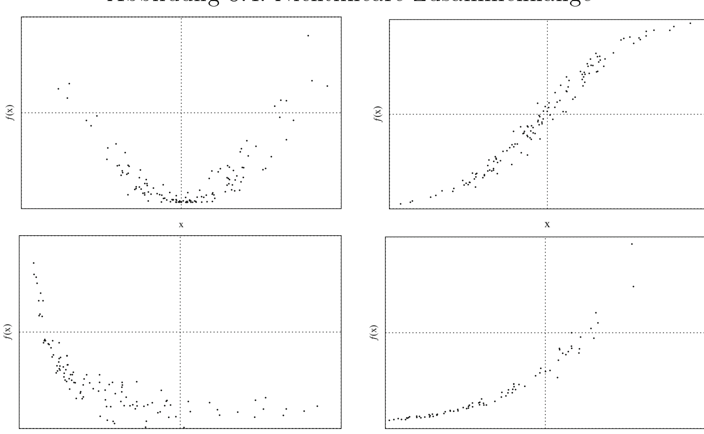
Nichtlineare Beziehungen

Nimmt R^2 sehr niedrige Werte an, so kann dies unterschiedliche Ursachen haben. Eine Möglichkeit besteht darin, dass die unabhängige Variable X tatsächlich keinen Einfluss auf Y ausübt, wie in Abbildung 7.7 auf Seite 174 ($r = 0$). Es könnte aber auch sein, dass ein *nichtlinearer* Zusammenhang besteht. In Abbildung 8.4 auf Seite 190 sind verschiedene nichtlineare Beziehungen dargestellt. Bei nichtlinearen Beziehungen ist der Anstieg in Y von der Position auf der X-Achse abhängig.

Tabelle 8.1: Berechnung des Determinationskoeffizienten R^2

Wahlkreis	x_i	y_i	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
Neuwied	55,55	44,21	0,25	0,06	44,07	0,14	0,02	0,11	0,01
Ahrweiler	81,99	50,13	6,17	38,07	49,20	0,93	0,86	5,24	27,46
Koblenz	73,14	46,60	2,64	6,97	47,48	−0,88	0,77	3,52	12,39
Cochem	70,78	50,94	6,98	48,72	47,02	3,92	15,37	3,06	9,36
Bad Kreuznach	32,60	39,10	−4,86	23,62	39,61	−0,51	0,26	−4,35	18,92
Bitburg	91,40	52,68	8,72	76,04	51,02	1,66	2,76	7,06	49,84
Trier	87,97	44,82	0,86	0,74	50,36	−5,54	30,69	6,40	40,96
Montabaur	50,76	43,42	−0,54	0,29	43,14	0,28	0,08	−0,82	0,67
Mainz	51,36	40,86	−3,10	9,61	43,25	−2,39	5,71	−0,71	0,50
Worms	32,81	37,99	−5,97	35,64	39,66	−1,67	2,79	−4,30	18,49
Frankenthal	31,98	39,71	−4,25	18,06	39,49	0,22	0,05	−4,47	19,98
Ludwigshafen	38,01	40,86	−3,10	9,61	40,66	0,20	0,04	−3,30	10,89
Neustadt-Speyer	45,61	46,48	2,52	6,35	42,14	4,34	18,84	−1,82	3,31
Kaiserslautern	34,89	37,68	−6,28	39,44	40,06	−2,38	5,66	−3,90	15,21
Pirmasens	45,98	42,79	−1,17	1,37	42,21	0,58	0,34	−1,75	3,06
Südpfalz	55,07	45,09	1,13	1,28	43,97	1,12	1,25	0,01	0,00
	$\bar{\mathbf{x}}$ = **54,99**	$\bar{\mathbf{y}}$ = **43,96**		$\mathbf{SAQ_y}$ = **315,87**			**U. $\mathbf{SAQ_y}$** = **85,49**		**E. $\mathbf{SAQ_y}$** = **231,05**

Abbildung 8.4: Nichtlineare Zusammenhänge



Nichtlineare Beziehungen können im Rahmen des linearen Regressionsmodells berücksichtigt werden, wenn durch eine geeignete Veränderung der unabhängigen oder abhängigen Variablen eine lineare Beziehung zwischen den transformierten Variablen hergestellt werden kann. Ein Beispiel: In der Ökonomie wird ein exponentieller Anstieg der Lohnhöhe mit zunehmender Zahl der Schuljahre angenommen: $\widehat{\texttt{Lohn}} = e^{a+b\cdot\texttt{Schuljahre}}$ (vgl. die rechte, untere Grafik in Abbildung 8.4). Durch Logarithmieren erhält man einen linearen Zusammenhang zwischen dem logarithmierten Lohn und der Zahl der Schuljahre: $\widehat{\ln(\texttt{Lohn})} = a + b \cdot \texttt{Schuljahre}$. Regressionskonstante und Regressionskoeffizient lassen sich dann mit einer linearen Regression schätzen. Beachtet werden muss allerdings, dass bei einem Anstieg der unabhängigen Variablen um eine Einheit nun der Logarithmus der abhängigen Variablen – $\ln y$ – um b Einheiten (linear) ansteigt (siehe Wooldridge 2006, 46 ff.).

Korrelation und Kausalität

Abschließend ein Hinweis zur Interpretation: Wir interpretieren die Ergebnisse von Regressionsgleichungen häufig *kausal*. X wird dann als ursächlich für Y angesehen. Wir haben in Kapitel 2.3 jedoch gesehen, dass der Nachweis von Kausalität bei nicht experimentellen Daten schwierig ist und

den Ausschluss alternativer Erklärungen über eine Drittvariablenkontrolle erfordert. Die Drittvariablenkontrolle erfolgt im Rahmen der Regressionsanalyse, indem die als relevant erachteten Drittvariablen als weitere unabhängige Variablen im Regressionsmodell aufgenommen werden. Mit einer *multiplen* Regression kann der Einfluss mehrerer unabhängiger Merkmale auf ein abhängiges Merkmal geschätzt werden. Die Interpretation der Koeffizienten erfolgt analog zur bivariaten Regression.

Regressionsanalysen sind auch dann sinnvoll, wenn wir ausschließlich an einer Prognose des abhängigen Merkmals interessiert sind. In Kapitel 7.6 wurde für die Lese- und Mathematikkenntnisse eine Produkt-Moment-Korrelation von $r = 0{,}88$ festgestellt. Die Regressionsgleichung für den in Abbildung 7.6 (S. 173) dargestellten Zusammenhang (Daten: IALS 1994, n=2062) wurde mit einem Statistik-Programm berechnet:

$$\widehat{\text{Mathe}} = 61,6 + 0,84 \cdot \text{Lese} \tag{8.13}$$

Für einen Anstieg der Lesekompetenz um einen Punkt prognostizieren wir einen Anstieg der Mathematikkompetenz um 0,8 Punkte. R^2 beträgt 77 %. Die Lesekompetenz ist damit sehr gut zur Prognose der Mathematikkompetenz geeignet. Kausal würden wir diesen Effekt der Lesekompetenz wohl nicht interpretieren. Ursächlich für die Mathematik- *und* die Lesekenntnisse sind andere Faktoren wie die Intelligenz oder die formale Bildung.

Die lineare Regression setzt metrische *abhängige Merkmale* voraus. In den Sozialwissenschaften haben wir jedoch häufig keine metrischen Merkmale. Etwa dann, wenn wir die Wahlabsicht von Befragten (nominal), die Erwerbstätigkeit von Frauen (dichotom) oder die Stärke des politischen Interesses (ordinal) erklären möchten. Auch für kategoriale Daten existieren Regressionsmodelle. Eine leicht verständliche Einführung findet sich bei Andreß et al. (1997, Kapitel 5).

Aufgaben zu Linearer Regression

1. Welche Fragestellungen können mit Hilfe der Regression beantwortet werden? (Bitte beantworten Sie die Frage in maximal 2 Sätzen).
2. Sie möchten wissen, welchen Einfluss der Anteil der Katholiken auf das Wahlergebnis der SPD bei der Bundestagswahl 1994 in Rheinland-Pfalz hatte. In der Tabelle sind für jeden rheinland-pfälzischen Wahlkreis der Anteil der Katholiken x_i und das Wahlergebnis der SPD y_i wiedergegeben.
   Bitte berechnen Sie die Regressionsgerade! Ist die ermittelte Regressionsfunktion eine gute Schätzung des Wahlergebnisses der SPD? Berechnen Sie zur Beantwortung dieser Frage das Bestimmtheitsmaß R^2! Interpretieren Sie alle errechneten Maße inhaltlich!

Wahlkreis	x_i	y_i
Neuwied	55,55	40,96
Ahrweiler	81,99	34,17
Koblenz	73,14	37,93
Cochem	70,78	32,84
Kreuznach	32,60	44,01
Bitburg	91,40	32,72
Trier	87,97	39,60
Montabaur	50,76	42,21
Mainz	51,36	36,55
Worms	32,81	42,42
Frankenthal	31,98	43,16
Ludwigshafen	38,01	40,83
Neustadt-Speyer	45,61	34,59
Kaiserslautern	34,89	46,70
Pirmasens	45,98	41,66
Südpfalz	55,07	36,93

3. Berechnen Sie für das obige Beispiel den Korrelationskoeffizienten aus den einzelnen Messwerten und aus dem Bestimmtheitsmaß!

9 Stichprobenziehung

Um Angaben über die Struktur der Bevölkerung zu erhalten, werden von amtlicher Seite regelmäßig Volkszählungen durchgeführt. Die letzte Volkszählung ist schon eine Weile her. Sie fand in der Bundesrepublik im Jahr 1987 und in der DDR im Jahr 1981 statt. Erhebt man die Daten wie im Falle von Volkszählungen bei allen interessierenden Untersuchungseinheiten – hier also bei der gesamten Bevölkerung eines Staates –, dann spricht man von einer *Vollerhebung*. Eine bevölkerungsweite Erhebung wie die Volkszählung ermöglicht unter anderem fein gegliederte regionale Analysen. Die Kehrseite der Medaille ist allerdings, dass bevölkerungsweite Vollerhebungen sehr kosten- und zeitintensiv sind. Nach Angaben von Diekmann (2008, 375) kostete die Volkszählung 1987 mehr als 1 Milliarde DM. Die Ausgaben für eine neue Volkszählung in Deutschland werden noch weitaus höher veranschlagt. Das Statistische Bundesamt schätzt die Kosten auf ca. 1,45 Mrd. Euro, das Deutsche Institut für Wirtschaftsforschung auf ca. 1 Mrd. Euro (Bundesministerium des Innern 2006; Wagner 2006). Aus Kostengründen wird der nächste Zensus im Jahr 2011 daher keine traditionelle Volkszählung mehr sein. Stattdessen wird ein *registergestützter Zensus* durchgeführt (Krügener 2006), dessen Kosten das Statistische Bundesamt mit ca. 450 Millionen Euro beziffert. Registergestützt heißt, dass Daten aus den Verwaltungsregistern genutzt werden – insbesondere aus den Melderegistern und den Registern der Bundesagentur für Arbeit. Die Register enthalten aber nicht alle interessierenden Merkmale. Angaben zur Bildung und Ausbildung sind in den Verwaltungsregistern beispielsweise nicht enthalten. Zusätzlich werden deshalb Umfragen bei einem Teil der Bürger und Bürgerinnen durchgeführt. Der Preis für den Verzicht auf eine traditionelle Vollerhebung besteht darin, dass man die Verteilung bestimmter Merkmale in der Grundgesamtheit nicht mehr kennt, sondern schätzen muss.

Zur Beantwortung wissenschaftlicher Fragestellungen sind bevölkerungsweite Vollerhebungen nicht nur aus Kostengründen ungeeignet. Würde

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen*,
DOI 10.1007/978-3-531-91879-2_9,

man alle wahlberechtigten Bundesbürger – also ca. 60 Millionen Menschen – vor einer Wahl nach ihrer Wahlabsicht befragen, dann lägen die Ergebnisse sicher nicht mehr vor der Wahl vor. Aus diesem Grund befragt man nicht alle Wahlberechtigten, sondern trifft eine *Auswahl*. Auswahlen werden auch als *Stichproben* oder *Samples* bezeichnet (vgl. zu Auswahlverfahren Scheaffer et al. 1996; Levy und Lemeshow 1991; Böltken 1976). In der nachfolgenden Tabelle ist die Wahlabsicht von 1.250 Befragten vor der Bundestagswahl 1994 und das tatsächliche Endergebnis wiedergegeben. Obwohl lediglich ein kleiner Teil der Wähler befragt wurde, weicht das Ergebnis der Umfrage vom tatsächlichen Wahlergebnis nicht sehr weit ab. Das ist aber nicht immer so. Bei der Bundestagswahl 2005 wurde der Stimmenanteil von CDU/CSU von nahezu allen Umfrageinstituten deutlich zu hoch ausgewiesen.

Tabelle 9.1: Umfrageergebnis und tatsächliches Ergebnis der BTW 1994

Partei	Forschungs-gruppe Wahlen	Amtliches Ergebnis
CDU/CSU	42,5 %	41,5 %
SPD	35,5 %	36,4 %
FDP	7,0 %	6,9 %
Bündnis 90/Grüne	8,0 %	7,3 %
PDS	3,5 %	4,4 %
REP	2,0 %	1,9 %
	1.250 (Befragte)	47.104.576 (Wähler)

Es ist kein Zufall, dass sich die Methoden der Stichprobenziehung parallel zur Wahlforschung entwickelten. Normalerweise können die Ergebnisse einer Stichprobe nicht an den Ergebnissen für die Grundgesamtheit validiert werden. Bei Wahlen ist das anders. Hier gibt das Wahlergebnis die Stimmenanteile für die einzelnen Parteien in der Grundgesamtheit an. Waren die Umfrageergebnisse weit vom tatsächlichen Wahlergebnis entfernt, so lag dies häufig an verzerrten Stichproben. „Fehlprognosen“ wurden zum Anlass genommen, die Stichprobenziehung und die Datenerhebung zu überdenken.

Die entscheidende Frage ist, wie man auf der Basis von Stichproben Aussagen über die Grundgesamtheit treffen kann. Diese Frage kann mit Hilfe

der *schließenden Statistik* bzw. *Inferenzstatistik* beantwortet werden. Im Gegensatz zur bisher behandelten *deskriptiven Statistik*, die sich der Beschreibung vorliegender Daten zufriedengibt, werden mit der schließenden Statistik Stichprobenergebnisse verallgemeinert.

Ob Aussagen über die Grundgesamtheit zulässig sind, hängt wesentlich davon ab, auf welche *Weise* die Einheiten der Stichprobe ausgewählt werden. Prinzipiell unterscheidet man zwischen *zufallsgesteuerten* und *nicht-zufallsgesteuerten* Auswahlverfahren. Schlüsse von der Stichprobe auf die Grundgesamtheit – im Beispiel also von 1.250 Befragten auf 47 Millionen Wähler – sind nur bei Zufallsstichproben theoretisch begründbar. Warum dies so ist und welche Möglichkeiten es gibt, Stichprobenresultate zu verallgemeinern, damit beschäftigen sich alle folgenden Kapitel.

9.1 Grundlagen

Bevor die verschiedenen Formen der Stichprobenziehung erläutert werden können, sollten einige Begriffe geklärt sein.

9.1.1 Grundgesamtheit, Auswahlgesamtheit und Stichprobe

Unter **Grundgesamtheit** oder **Population** werden alle Einheiten verstanden, auf die sich die Untersuchungshypothesen beziehen, wobei die Einheiten real existieren müssen (vgl. zur Annahme fiktiver Grundgesamtheiten die Kritik von Rohwer und Pötter 2002). Interessiert man sich für das Wahlverhalten der Deutschen bei einer Bundestagswahl, dann stellen alle bei dieser Wahl wahlberechtigten Bundesbürger die Grundgesamtheit dar. Soll die Wahlkampfberichterstattung der auflagenstärksten überregionalen Tageszeitungen (ohne Boulevardblätter) bei der Bundestagswahl 2009 inhaltsanalytisch ausgewertet werden, dann zählen alle wahlkampfbezogenen Artikel der „Frankfurter Allgemeinen Zeitung“, der „Frankfurter Rundschau“, der „Süddeutschen Zeitung“, der „tageszeitung“ und der „Welt“ zur Grundgesamtheit. Sollen die Studienwünsche rheinland-pfälzischer Abiturienten untersucht werden, dann gehören alle Schüler des 13. Schuljahres in Rheinland-Pfalz zur Grundgesamtheit. Kennwerte der Grundgesamtheit (bzw. genauer: einer theoretischen Verteilung) werden als Parameter bezeichnet. Zur Darstellung werden griechische Buchstaben verwandt – z. B. kennzeichnet μ (sprich: mü) das arithmetische Mittel in der Grundgesamtheit und σ^2 (sprich: sigma) die Varianz.

Von der Grundgesamtheit ist die **Auswahlgesamtheit** zu unterscheiden. Sie besteht aus allen Einheiten, aus denen die Stichprobe tatsächlich ausgewählt wird. Zur Untersuchung der Studienwünsche rheinland-pfälzischer Abiturienten könnten wir beispielsweise alle Gymnasien anschreiben und deren Direktoren bitten, uns eine Liste aller Schüler des 13. Schuljahres zu schicken. Die Auswahlgesamtheit besteht dann aus den auf diesen Listen verzeichneten Schülern, die Grundgesamtheit aus allen rheinland-pfälzischen Schülern. Die Auswahlgesamtheit und nicht die Grundgesamtheit ist demnach die *Grundlage der Stichprobenziehung*.

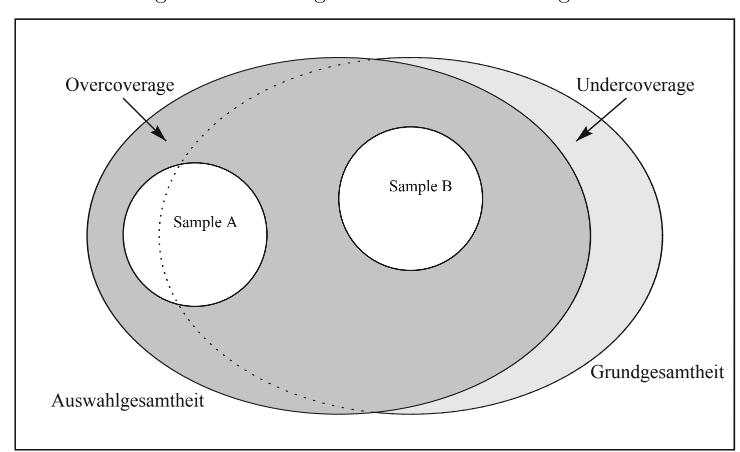
Die Auswahlgesamtheit kann sich von der Grundgesamtheit durch **undercoverage** (Untererfassung) und **overcoverage** (Übererfassung) unterscheiden. Undercoverage liegt dann vor, wenn Einheiten der Grundgesamtheit nicht in der Auswahlgesamtheit vorhanden sind; overcoverage, wenn Einheiten der Auswahlgesamtheit nicht zur Grundgesamtheit gehören. Fehlen Schüler des 13. Jahrgangs auf unserer Liste, z. B. weil diese erst nach Erstellung der Liste aus einem anderen Bundesland zugezogen sind, dann liegt undercoverage vor. Enthält unsere Liste andererseits Schüler, die zwischenzeitlich das Gymnasium verlassen haben, besteht overcoverage. Die Differenz zwischen Auswahl- und Grundgesamtheit wäre in diesen beiden Fällen durch eine veraltete Liste verursacht.

Eine **Stichprobe** ist eine Teilmenge von Untersuchungseinheiten, die nach bestimmten Regeln, dem Auswahlplan, ausgezählt wurde. Synonym verwendet man auch die Begriffe *Auswahl* oder *Sample*. Wir könnten z. B. bei der Auswahl von Schülern in Rheinland-Pfalz so vorgehen, dass wir erst eine Schülerliste erstellen und dann aus dieser Schülerliste jeden zehnten Schüler auswählen. Kennwerte von Stichproben werden häufig als Statistiken (*sample statistic*) bezeichnet. Wir haben schon diverse Statistiken für Stichproben kennen gelernt, z. B. $\bar{x}$ für das arithmetische Mittel oder s^2 für die Varianz.

Der Zusammenhang zwischen Grund- und Auswahlgesamtheit ist in der folgenden Abbildung verdeutlicht: Die Grundgesamtheit wird durch die helle Ellipse dargestellt, die Auswahlgesamtheit durch die darüberliegende dunkle Ellipse. Die hellen Kreise sind zwei verschiedene Stichproben A und B. Under- und overcoverage sind die beiden sichelförmigen Überlappungen rechts und links. Wie man sieht, enthält Stichprobe A Einheiten, die nicht zur Grundgesamtheit gehören (was über die gestrichelte Linie hinausgeht). Dagegen sind in Stichprobe B nur Fälle verzeichnet, die so-

wohl zur Auswahl- als auch zur Grundgesamtheit gehören. Da jede Stichprobe auf der Auswahlgesamtheit beruht, können Fälle, die zwar in der Grundgesamtheit, nicht aber in der Auswahlgesamtheit vorhanden sind (undercoverage), nie in eine Stichprobe gelangen. Je weniger Auswahl- und Grundgesamtheit voneinander abweichen, umso besser die Grundlage der Stichprobenziehung. Sprachlich vernachlässigt man den Unterschied zwischen Auswahl- und Grundgesamtheit häufig, auch wenn klar ist, dass die Auswahlgesamtheit und nicht die Grundgesamtheit der Stichprobenziehung zugrunde liegt.

Abbildung 9.1: Auswahlgesamtheit und Grundgesamtheit



Bei der Stichprobenziehung muss zwischen der *Auswahleinheit* (Erhebungseinheit) und der *Untersuchungseinheit* unterschieden werden. Die Auswahleinheit ist die Einheit, die der Stichprobenziehung zugrunde liegt; die Untersuchungseinheit ist die Einheit, die Merkmalsträger ist. Im obigen Beispiel waren Auswahl- und Untersuchungseinheit identisch, nämlich Schüler. Genauso gut könnten wir auch Schulen auswählen und jeden Schüler der ausgewählten Schulen befragen. Die Auswahleinheit sind nun Schulen, die Untersuchungseinheit sind weiterhin die Schüler; schließlich interessieren wir uns für deren Studienwünsche.

9.1.2 Befragungsverweigerung

Angestrebt wird, die Daten für die gesamte Stichprobe zu erheben. Bei den meisten Befragungen treten jedoch in erheblichem Umfang Ausfälle auf (vgl. Schnell et al. 2008, 289). Gelingt es nicht, eine Person zu befragen, dann sprechen wir von emphUnit-Nonresponse. Die Nichtbeantwortung einzelner Fragen (*Item-Nonresponse*) wurde bereits in Kapitel 4 angesprochen.

Unproblematisch sind **Ausfälle**, die keine Auswirkung auf die Qualität der Stichprobe haben. Solche Ausfälle werden als *stichprobenneutral* bzw. *zufällig* bezeichnet. Das Wort zufällig deutet an, dass die realisierte Stichprobe dann als Zufallsstichprobe aus der angestrebten Stichprobe angesehen wird. Stichprobenneutralität wird beispielsweise unterstellt, wenn eine Zielperson nicht befragt werden konnte, weil deren Adresse nicht richtig notiert wurde. Zu einer Minderung der Stichprobenqualität führen dagegen nicht stichprobenneutrale bzw. *systematische Ausfälle* (*Non-Response Error*). Diese werden durch das Untersuchungsthema oder das Untersuchungsdesign verursacht. Ein systematischer Ausfall läge z. B. dann vor, wenn im Haushalt nie jemand angetroffen wird, weil alle Haushaltsangehörigen berufstätig sind, oder Berufstätige häufiger aus Zeitgründen das Interview verweigern. Ein systematischer Ausfall bestände auch dann, wenn vor allem Befragte mit rechtsextremen Einstellungen eine Befragung zum Thema Rechtsextremismus ablehnen. In beiden Fällen wären in der Stichprobe bestimmte Gruppen – Berufstätige bzw. Befragte mit rechtsextremen politischen Einstellungen – im Vergleich zur Grund- bzw. Auswahlgesamtheit unterrepräsentiert. Die Stichproben würden systematisch von der Grundgesamtheit abweichen.

Ob ein Ausfall stichprobenneutral ist oder nicht, lässt sich nur schwer feststellen. Bei mündlichen und telefonischen Befragungen können die Interviewer nachfragen, warum eine Person die Teilnahme an der Untersuchung ablehnt. In Tabelle 9.2 auf der gegenüberliegenden Seite sind die Ausfälle von Befragungspersonen beim ALLBUS 2006 wiedergegeben. Wie man sieht, gibt es nur sehr wenige stichprobenneutrale Ausfälle, zu denen Nicht-Befragung aufgrund falscher Adressen, Wohnungswechsel und Tod gezählt werden. Personen, die nicht in Privathaushalten wohnen, zählen qua Definition nicht zur Grundgesamtheit. Beim ALLBUS 2006 belief sich der Anteil stichprobenneutraler Ausfälle an der Bruttostichprobe auf 11 %

in West- und 9,2 % Ostdeutschland. Die systematischen Ausfälle resultieren zum größten Teil aus *Befragungsverweigerungen*. Diese werden zu den systematischen Ausfällen gezählt, weil man davon ausgeht, dass sich die Personen, die eine Teilnahme verweigern, von den teilnahmebereiten Personen unterscheiden (vgl. Diekmann 2008, 422). Ein systematischer Ausfall liegt natürlich auch dann vor, wenn eine Person nicht in der Lage ist, ein Interview in deutscher Sprache zu führen. Diese Ausfälle führen zu einer Unterrepräsentation von Migranten in der Stichprobe. Migranten, die Fragen in deutscher Sprache beantworten können, unterscheiden sich sehr wahrscheinlich auch in anderen Merkmalen von Migranten, die dazu nicht in der Lage sind (z. B. im Bildungsniveau).

Tabelle 9.2: Ausschöpfung beim ALLBUS 2006

		West		**Ost**	
		n	%	n	%
Ursprüngliche Bruttostichprobe		5772	100	2652	100
Zusätzlich eingesetzte Adressen als Ersatz für stichprobenneutrale Ausfälle	+	647	11,2	232	8,7
Bruttostichprobe	=	6419	100	2884	100
Stichprobenneutrale Ausfälle insgesamt	−	704	11,0	264	9,2
– Anschreiben nicht zustellbar		132	2,1	55	1,9
– Adresse falsch, existiert nicht (mehr)		122	1,9	41	1,4
– ZP verstorben		40	0,6	15	0,5
– ZP verzogen		343	5,3	133	4,6
– ZP lebt nicht in Privathaushalt		67	1,0	20	0,7
Bereinigte Bruttostichprobe	=	5715	100	2620	100
Systematische Ausfälle insgesamt	−	3416	59,8	1498	57,2
– Im Haushalt niemand angetroffen		238	4,2	93	3,5
– ZP nicht angetroffen		137	2,4	63	2,4
– ZP nicht befragungsfähig		167	2,9	86	3,3
– ZP verweigert tel. bei Infratest Projektleitung		28	0,5	26	1,0
– ZP aus Zeitgründen nicht zum Interview bereit		261	4,6	107	4,3
– ZP generell nicht zum Interview bereit		2366	41,4	1080	41,2
– ZP spricht nicht hinreichend genug Deutsch		121	2,1	10	0,4
– Adresse nicht abschließend bearbeitet		26	0,5	15	0,6
– Interviews als (Teil-)Fälschung identifiziert		72	1,3	18	0,7
Auswertbare Interviews	=	**2299**	**40,2**	**1122**	**42,8**

Quelle: Wasmer et al. (2007), S. 68.

Um die Qualität der Stichprobe angeben zu können, wird häufig die *Ausschöpfungsquote* berechnet. Sie bezeichnet den Anteil realisierter Inter-

views an einer „bereinigten Bruttostichprobe". Die „bereinigte Bruttostichprobe" ist die um die stichprobenneutralen Ausfälle bereinigte Zahl aller zu befragenden Personen. Wie hoch die Ausschöpfungsquote ist, hängt also maßgeblich davon ab, was als stichprobenneutraler Ausfall gezählt wird (vgl. Koch 1993). Bei seriösen Studien werden deshalb neben der Ausschöpfungsquote auch die Art der Ausfälle angegeben (vgl. für das SOEP Hanefeld 1987, 182, 184). Die Ausschöpfungsquote betrug beim ALLBUS 2006 in Westdeutschland 40,2 % und in Ostdeutschland 42,8 %. Im ALLBUS 1994 lag die Ausschöpfungsquote noch bei 53,2 % (West) bzw. 55,2 % (Ost) (Koch et al. 1994). Die Ausschöpfungsquote liegt beim ALLBUS 2006 also rund 10 Prozentpunkte niedriger als beim ALLBUS 2004. Die Entwicklung beim ALLBUS zeigt den allgemeinen Trend hin zu höherer Befragungsverweigerung.

Der Umfang des Unit-Nonresponse lässt sich bei allen Formen der Befragung durch eine Erhöhung der Kontaktversuche – also wiederholtes Anschreiben, Antelefonieren, mehrmalige Interviewerbesuche –, kleine Geschenke (*incentives*), eine spezielle Schulung der Interviewer usw. reduzieren (vgl. Dillman 1978). Im *European Social Survey* wird eine sehr ambitionierte Ausschöpfungsquote (*response rate*) von 70 % in den teilnehmenden Staaten angestrebt. In der ersten Runde (2002/2003) wurde diese zwar in einer Reihe von Staaten (zum Teil deutlich) unterschritten. In den Niederlanden, deren Bevölkerung eher als ‚befragungsmüde' gilt, wurde allerdings eine Ausschöpfungsquote realisiert, die nur geringfügig unter der Zielmarke von 70 % lag und deutlich über der nationaler Erhebungen (Billiet et al. 2007). Die Zahl der Kontaktversuche und ein spezielles Interviewertraining zur Konversion von Befragungsverweigerern wird für die hohe Ausschöpfungsquote verantwortlich gemacht.

Ausfälle werden sich aber auch bei einer sorgfältigen Datenerhebung kaum vermeiden lassen. Umfrageinstitute versuchen, das Problem systematischer Ausfälle durch die Konstruktion von Gewichtungsfaktoren zu beheben (vgl. Gabler et al. 1994; Elliot 1991). Gruppen, die in der Stichprobe im Vergleich zur Grundgesamtheit unterrepräsentiert sind, werden bei der Datenanalyse höher gewichtet und Gruppen, die in der Stichprobe im Vergleich zur Grundgesamtheit überrepräsentiert sind, werden niedriger gewichtet, und zwar so, dass die Anteile in der Stichprobe denen der Grundgesamtheit entsprechen. Technisch ist diese Art der Gewichtung, die als *Redressment* (Nachgewichtung) bezeichnet wird, mit Statistikprogrammen leicht zu realisieren.

Gewichtungsfaktoren können nur für Merkmale gebildet werden, deren Verteilung in der Grundgesamtheit bekannt ist, etwa durch Volkszählungen. Nur für diese Merkmale lassen sich auch systematische Abweichungen von der Grundgesamtheit feststellen. Nicht kontrollierbar ist allerdings, ob die Ausfälle innerhalb einer Gruppe rein zufällig erfolgt sind: Werden die Antworten von Migranten hochgewichtet, weil Migranten in der realisierten Stichprobe unterrepräsentiert sind, dann fußt dies auf der Annahme, dass sich das Antwortverhalten der befragten Migranten nicht von dem Antwortverhalten der Migranten unterscheidet, die nicht an der Befragung teilgenommen haben. Zudem bleibt Item-Nonresponse bei der Gewichtung unberücksichtigt.

Bei Datenanalysen werden in der Regel die Untersuchungseinheiten berücksichtigt, die auf den interessierenden Merkmalen keinen einzigen fehlenden Wert aufweisen. Soll der Einfluss des Alters, des Geschlechts und der Bildung auf das Erwerbseinkommen mit einer *multiplen Regression* untersucht werden, dann werden ausschließlich die Personen analysiert, für die Angaben zu allen vier Merkmalen vorhanden sind. Diese Methode wird als *listwise deletion* (listenweiser Fallausschluss) bezeichnet. Gerechtfertigt ist diese Vorgehensweise nur dann, wenn die kompletten Fälle als eine Zufallsstichprobe aus allen Fällen aufgefasst werden können (keine systematischen Ausfälle). Diese Annahme ist in vielen Fällen nicht haltbar. In den vergangenen Jahren wurde daher eine Reihe von statistischen Verfahren zur Behandlung von fehlenden Werten (*missing values*) (weiter-)entwickelt, die geringere Anforderungen an den Ausfallmechanismus stellen, weil sie alle beobachteten Informationen ausnutzen (vgl. dazu Allison 2002; Little und Rubin 2002).

9.2 Zufall und Wahrscheinlichkeit

Wenn im Alltag von „Zufall" gesprochen wird, meint man damit meist ein willkürliches Ereignis, das keiner bestimmten und nachvollziehbaren Gesetzmäßigkeit unterliegt. Im mathematischen Sinne wird ein Ereignis als zufällig bezeichnet, wenn es das Resultat eines Zufallsexperiments ist. Bei einem Zufallsexperiment sind die *möglichen* Ereignisse bekannt, nicht aber welches Ereignis tatsächlich eintritt. Zufallsexperimente sind zumindest theoretisch unendlich häufig wiederholbar. Beispiele für Zufallsexperimente sind das Werfen einer Münze oder eines Würfels, das Ziehen der Lottozahlen oder die Ziehung einer Zufallsstichprobe. Das Auftreten der

Ereignisse eines Zufallsexperiments ist mit Wahrscheinlichkeiten mathematisch exakt beschreibbar.

Ein einfaches und immer wieder gern benutztes Beispiel ist das Werfen eines normalen Würfels. „Normal" soll heißen, dass der Würfel sechs gleiche Seiten hat. Die möglichen Ergebnisse des Wurfs lassen sich mit den Ziffern bezeichnen, die auf dem Würfel angegeben sind. Die Menge der Elementarergnisse besteht aus den Zahlen 1, 2, 3, 4, 5 und 6. Welches dieser Elementarereignisse auftritt, ist dem Zufallsprozess überlassen. Jedes dieser Ereignisse hat die gleiche Auftretenswahrscheinlichkeit, wenn der Würfel nicht manipuliert ist und man beim Werfen nicht schummelt.

Wahrscheinlichkeiten können mit Zahlen im Bereich von 0 bis 1 oder mit entsprechenden Prozentwerten (0 bis 100 %) bezeichnet werden. Ein sicheres Ereignis hat die Wahrscheinlichkeit 1 bzw. 100 %, ein unmögliches Ereignis die Wahrscheinlichkeit 0 bzw. 0 %. Die Summe der Wahrscheinlichkeiten aller Elementarereignisse ist 1, weil sich Elementarereignisse gegenseitig ausschließen. Daraus folgt, dass die Gegenwahrscheinlichkeit eines Ereignisses 1 abzüglich der Wahrscheinlichkeit des Ereignisses ist. Da beim Werfen des Würfels irgendeine Zahl fallen muss – wir schließen also aus, dass der Würfel auf der Kante stehenbleiben könnte – und nicht zwei Ziffern gleichzeitig auftreten können, lässt sich sagen, dass mit 100%iger Wahrscheinlichkeit eine Zahl zwischen 1 und 6 fallen wird.

Sind alle Elementarereignisse wie beim Werfen eines Würfels gleich wahrscheinlich, dann lassen sich die Wahrscheinlichkeiten P der Elementarereignisse $j = 1, \cdots, n$ mit

$$P(\{j\}) = \frac{1}{N}$$

berechnen, wobei N hier die Zahl der möglichen Elementarereignisse bezeichnet. Für den Wurf eines Würfels ermittelt man für jedes Elementarereignis die Wahrscheinlichkeit $P = 1/6 = 0{,}1\bar{6}$.

Sind die Elementarereignisse eines Zufallsexperiments gleich wahrscheinlich (*Laplace*-Experiment), dann lässt sich die Wahrscheinlichkeit aller möglichen Ereignisse eines Zufallsexperiments durch Zählen der günstigen Ereignisse A (d. h. das Eintreten von A) im Vergleich zu allen möglichen Ereignissen ermitteln:

$$P(A) = \frac{\text{Zahl der für A günstigen Ereignisse}}{\text{Zahl aller möglichen Ereignisse}} \,. \tag{9.1}$$

Diese Wahrscheinlichkeit wird auch als Laplace-Wahrscheinlichkeit oder *a priori*-Wahrscheinlichkeit bezeichnet. *A priori*, weil die Wahrscheinlichkeit $P(A)$ *vor* der Durchführung des Zufallsexperiments theoretisch bestimmt werden kann.

Ein Beispiel: Man könnte danach fragen, wie wahrscheinlich das Werfen einer *geraden Zahl* ist. Die möglichen günstigen Ereignisse werden mit „oder" verknüpft: „Wie wahrscheinlich ist der Wurf einer 2 oder einer 4 oder einer 6?" Drei von sechs möglichen Elementarereignissen sind gerade Zahlen:

$$P(\text{gerade Zahl}) = P(2\,\text{oder}\,4\,\text{oder}\,6) = P(2 \cup 4 \cup 6) = \frac{3}{6} = \frac{1}{2} = 0{,}5 \,. \tag{9.2}$$

Dem entspricht die Addition der Einzelwahrscheinlichkeiten (Additionstheorem), weil sich die möglichen Ereignisse des Zufallsexperiments ‚einmaligen Werfens eines Würfels' gegenseitig ausschließen:

$$P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2} = 0{,}5 \,. \tag{9.3}$$

Die Wahrscheinlichkeit des Wurfs einer geraden Zahl beträgt also 0,5.

Man kann auch danach fragen, wie wahrscheinlich es ist, zweimal hintereinander eine 6 zu werfen, also eine 6 *und* noch einmal eine 6, $P(6 \cap 6)$. Insgesamt hat das Experiment zweimaliges Werfen eines Würfels 36 Elementarereignisse (Tabelle 9.3). Das günstige Ereignis ist $(6, 6)$.

Eines der 36 möglichen Ereignisse ist das zweimalige Werfen einer 6, also ist die Wahrscheinlichkeit $p(6 \cap 6) = 1/36$. Die Wahrscheinlichkeit lässt sich aus der Multiplikation der Einzelwahrscheinlichkeiten (Multiplikationstheorem für unabhängige Ereignisse) berechnen:

Tabelle 9.3: Mögliche Ereignisse beim zweimaligen Werfen eines Würfels

Ergebnis des 2. Wurfs	Ergebnis des 1. Wurfs: 1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

$$P(6) \times P(6) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 0{,}02\bar{7}\,. \tag{9.4}$$

Die Wahrscheinlichkeit, zweimal nacheinander eine 6 zu werfen, beträgt also 1/36 bzw. $0{,}02\bar{7}$.

Weil man beim Würfeln häufig an hohen Zahlen interessiert ist, könnte man auch fragen, mit welcher Wahrscheinlichkeit man zweimal hintereinander eine Zahl *größer als* 4 wirft. Diese Wahrscheinlichkeit lässt sich durch Auszählen leicht bestimmen (Tabelle 9.3). Sie beträgt $4/36 = 1/9 = 0{,}11$.

Wir werden im nächsten Kapitel noch einen anderen, den frequentistischen, Wahrscheinlichkeitsbegriff kennen lernen. Eine übersichtliche Einführung in die Wahrscheinlichkeitstheorie findet sich bei Kühnel und Krebs (2007, 107-127).

9.3 Zufallsgesteuerte Auswahlverfahren

Ein Auswahlverfahren wird als zufällig bezeichnet, wenn jede Einheit der Auswahlgesamtheit eine *gleiche bzw. eine angebbare Wahrscheinlichkeit*

größer null hat, in die Stichprobe zu gelangen. Die gleiche oder bekannte Chance wird durch Zufallsverfahren realisiert.

Wollen wir eine Stichprobe von 500 der 9.700 rheinland-pfälzischen Abiturienten (Angaben laut Statistischem Landesamt Rheinland-Pfalz für 1995) ziehen, und besitzt jeder Abiturient die *gleiche* Wahrscheinlichkeit in die Stichprobe zu gelangen, dann beträgt die Auswahlwahrscheinlichkeit $\frac{500}{9700} = 0{,}052$ (also ca. 5 %). Daraus folgt unmittelbar, dass Merkmalsausprägungen, die häufig in der Auswahlgesamtheit vorkommen, auch eine hohe Wahrscheinlichkeit haben, in die Stichprobe zu gelangen, und umgekehrt. Nehmen wir an, dass von den 9.700 rheinland-pfälzischen Abiturienten 400 Medizin und 3 Byzantistik studieren möchten. Die Wahrscheinlichkeit, einen Abiturienten mit Medizin-Studienwunsch auszuwählen, ist deutlich höher ($400 \cdot 1/9700 = 0{,}041$) als die Wahrscheinlichkeit, einen Abiturienten mit Byzantistik-Studienwunsch auszuwählen ($3 \cdot 1/9700 = 0{,}0003$).

9.3.1 Einfache Zufallsauswahlen

Bei einer einfachen (auch: uneingeschränkten) Zufallsauswahl (*simple random sample*) besitzt jede mögliche Stichprobe vom Umfang n aus der Grundgesamtheit (und jedes Element der Grundgesamtheit) dieselbe Auswahlwahrscheinlichkeit. Alle Stichprobenelemente werden unabhängig voneinander per Zufall in einem Auswahlvorgang ermittelt.

Technisch kann dies z. B. durch eine Lostrommel (Lotterieauswahl) oder durch Zufallszahlen geschehen (vgl. zum konkreten Verfahren Babbie 1997, 214 f.). Nehmen wir an, wir hätten eine Kartei, in der alle 9.700 rheinland-pfälzischen Abiturienten verzeichnet wären. Soll unsere Stichprobe 500 Schüler enthalten, dann können wir die Kartei durchnummerieren und 500 Zufallszahlen zwischen 1 und 9.700 erzeugen (jede Zahl hat die gleiche Wahrscheinlichkeit) bzw. einer Tabelle entnehmen. Diejenigen 500 Schüler, deren Nummern mit den Zufallszahlen übereinstimmen, gelangen in die Stichprobe. Genauso gut könnten wir die Namen der Schüler auf Zettel schreiben, diese in eine Lostrommel stecken und mischen. Aus dieser Trommel müssten dann nacheinander 500 Zettel gezogen und die Namen notiert werden. Jeder Schüler hat – wie wir bereits oben gesehen haben – eine Auswahlwahrscheinlichkeit von $500/9700 = 0{,}052$.

Für die Stichprobe ist es unerheblich, in welcher Reihenfolge die einzelnen Schüler gezogen werden. Zudem kann jeder Schüler nur einmal in die Auswahl gelangen – wir ziehen die Stichprobe *ohne Zurücklegen.* Aus einer Grundgesamtheit der Größe N können insgesamt $\binom{N}{n}$ (sprich: N über n) verschiedene Stichproben des Umfangs n ohne Berücksichtigung der Anordnung und ohne Zurücklegen gezogen werden. Der Ausdruck $\binom{N}{n}$ ist der Binomialkoeffizient und wird als „N über n" gelesen.

$$\binom{N}{n} = \frac{N!}{n! \cdot (N-n)!} \tag{9.5}$$

$n!$ ist die Fakultät von n, also $n \cdot (n-1) \cdot (n-2) \cdot 3 \cdot 2 \cdot 1$, $N!$ ist entsprechend $N \cdot (N-1) \cdot (N-2) \cdot 3 \cdot 2 \cdot 1$. Zur Illustration gehen wir von einer Grundgesamtheit von $N = 4$ Elementen aus, nämlich $\{\text{Vater}, \text{Opa}, \text{Mutter}, \text{Oma}\}$. Wie viele Stichproben der Größe $n = 2$ können aus dieser Grundgesamtheit gezogen werden? Insgesamt können

$$\binom{4}{2} = \frac{4!}{2! \cdot (4-2)!} = \frac{24}{4} = 6 \tag{9.6}$$

Stichproben vom Umfang $n = 2$ aus einer Grundgesamtheit von $N = 4$ Elementen gezogen werden. Die Zusammensetzung der einzelnen Stichproben ist $S_1 = (\text{Vater}, \text{Opa})$, $S_2 = (\text{Vater}, \text{Mutter})$, $S_3 = (\text{Vater}, \text{Oma})$, $S_4 = (\text{Opa}, \text{Mutter})$, $S_5 = (\text{Opa}, \text{Oma})$ und $S_6 = (\text{Mutter}, \text{Oma})$, wie in Tabelle 9.4 zu sehen ist. Die Wahrscheinlichkeit, dass eine bestimmte dieser Stichproben realisiert wird, beträgt $P(\text{Stichprobe}) = 1/\binom{4}{2} = 1/6$.

Aus einer Grundgesamtheit von 9.700 Schülern lassen sich schon mehr als 10^{853} verschiedene Stichproben vom Umfang 500 ziehen. Ein bekanntes Anwendungsbeispiel für Gleichung 9.5 ist die Frage nach der Wahrscheinlichkeit, 6 Richtige im Lotto 6 aus 49 zu erzielen. Die Anzahl der Möglichkeiten, 6 aus 49 Zahlen zu ziehen, beträgt $\binom{49}{6} = \frac{49!}{6! \cdot (49-6)!} = 13983816$. Die Wahrscheinlichkeit, dass *eine bestimmte Kombination* fällt (am besten natürlich die, die man selbst getippt hat) beträgt also $1/\binom{49}{6} = 1/13983816$. Die Ziehung der Lottozahlen entspricht dem Ziehen einer Stichprobe vom Umfang $n = 6$ aus einer Grundgesamtheit vom Umfang $N = 49$ ohne Berücksichtigung der Reihenfolge und ohne Zurücklegen.

Tabelle 9.4: Wahrscheinlichkeiten für Stichproben

	Stichprobe	p(Stichprobe)	Frauenanteil
S_1	Vater, Opa	1/6	0 %
S_2	Vater, Mutter	1/6	50 %
S_3	Vater, Oma	1/6	50 %
S_4	Opa, Mutter	1/6	50 %
S_5	Opa, Oma	1/6	50 %
S_5	Mutter, Oma	1/6	100 %

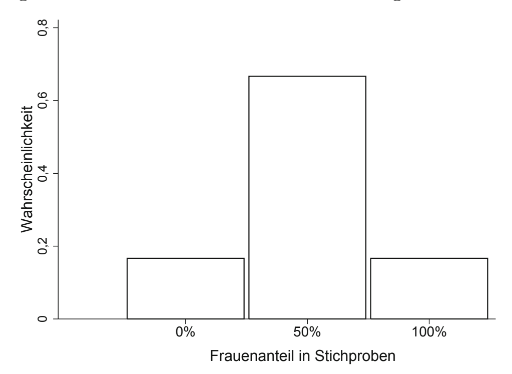
Zurück zum Beispiel, das in Tabelle 9.4 dargestellt ist. Aufgrund der unterschiedlichen Zusammensetzung der Stichproben variieren auch die Kennwerte (Mittelwerte, Anteilswerte etc.). In der rechten Spalte ist der Frauenanteil angegeben. In vier der sechs Stichproben beträgt der Frauenanteil 50 %, in einer Stichprobe (Vater, Opa) 0 % und in einer Stichprobe (Mutter, Oma) 100 %. In zwei Stichproben wird der Frauenanteil unter- bzw. überschätzt, in vier Stichproben stimmt der Frauenanteil mit dem Anteil in der Grundgesamtheit überein. Beim Frauenanteil in Stichproben handelt es sich um eine *Zufallsvariable* P . Bei Zufallsvariablen geben die Ausprägungen die Ereignisse eines Zufallsexperiments an. Im Beispiel ist das Zufallsexperiment das Ziehen einer Stichprobe. Die Zufallsvariable Frauenanteil hat die Ausprägungen 0 %, 50 % und 100 %.

Weil wir die Wahrscheinlichkeit kennen, mit der die einzelnen Stichproben gezogen werden (1/6), können wir auch angeben, mit welcher Wahrscheinlichkeit verschiedene Frauenanteile in Stichproben realisiert werden. Die Wahrscheinlichkeit einen Frauenanteil von 0 % zu erhalten, beträgt $1 \cdot 1/6$. In vier Stichproben beträgt der Frauenanteil 50 %. Die Wahrscheinlichkeit, in einer Stichprobe einen Frauenanteil von 50 % zu erhalten, beträgt daher $4 \cdot 1/6 = 4/6 = 2/3$. Die Wahrscheinlichkeit, einen Frauenanteil von 100 % zu erhalten, beträgt $1 \cdot 1/6 = 1/6$.

Die Wahrscheinlichkeitsverteilung des Frauenanteils ist in Abbildung 9.2 visualisiert. Auf der x-Achse sind die Frauenanteile abgetragen, auf der y-Achse deren Wahrscheinlichkeit.

Die Wahrscheinlichkeitsverteilung eines Stichprobenkennwerts – hier des Anteilswerts – wird als Kennwerteverteilung bezeichnet. Kennwerteverteilungen geben an, wie stark sich die Kennwerte in *möglichen* Stichproben

Abbildung 9.2: Wahrscheinlichkeitsverteilung des Frauenanteils



Stichproben vom Umfang n=2 aus einer Grundgesamtheit vom Umfang N=4 (ohne Berücksichtigung der Anordnung und ohne Zurücklegen).

von den Parametern der Grundgesamtheit unterscheiden. Als Stichprobenfehler wird die Abweichung der Stichprobenkennwerte vom Parameter der Grundgesamtheit bezeichnet. Der Stichprobenfehler hängt von zwei Faktoren ab: der Größe der Stichproben und der Streuung des interessierenden Merkmals in der Grundgesamtheit.

a) Je größer der Stichprobenumfang, umso schmaler ist die Kennwerteverteilung, d. h. Abweichungen der Stichprobenkennwerte vom Parameter der Grundgesamtheit werden kleiner. Bei einem Stichprobenumfang von $n = 2$ variiert der Frauenanteil in Stichproben zwischen 0 % und 100 % (Abbildung 9.2). In möglichen Stichproben des Umfangs $n = 3$ aus der Grundgesamtheit {Vater, Opa, Mutter, Oma} variiert der Frauenanteil zwischen 1/3 und 2/3. Im Extremfall besteht unsere Stichprobe aus allen Elementen der Grundgesamtheit. In diesem Fall muss der Kennwert der Stichprobe mit dem Parameter der Grundgesamtheit übereinstimmen.
b) Ein anderer Extremfall besteht, wenn das interessierende Merkmal in der Grundgesamtheit keine Streuung aufweist. Im Beispiel wäre dies der Fall, wenn unsere Grundgesamtheit aus vier Frauen bestände. Der Frauenanteil in der Grundgesamtheit wäre dann 100 %. In jeder der möglichen Stichproben aus dieser Grundgesamtheit beträgt der Frau-

enanteil dann ebenfalls 100 %. Je größer die Streuung eines Merkmals in der Grundgesamtheit, umso stärker streuen auch die Kennwerte in Stichproben.

Diese Angaben zum Stichprobenfehler werden im nächsten Kapitel präzisiert. Wichtig ist, dass Wahrscheinlichkeitsverteilungen für Stichprobenkennwerte die Verbindung zwischen dem (unbekannten) Parameter der Grundgesamtheit und den möglichen Stichprobenkennwerten herstellen. Die Zusammensetzung und die Kennwerte möglicher Stichproben sind Realisationen des Zufallsexperiments *Ziehen einer Stichprobe* und deshalb berechenbar.

Existieren zentrale Einwohnermelderegister wie in den skandinavischen Staaten, dann lassen sich problemlos einfache Zufallsstichproben für Bevölkerungsumfragen ziehen – vorausgesetzt das zentrale Einwohnermelderegister darf als Auswahlgrundlage genutzt werden. Für den *European Social Survey* wurden unter anderem in Dänemark, Norwegen, Schweden und Finnland einfache Zufallsstichproben aus den Einwohnermelderegistern gezogen (Häder und Lynn 2007).

Einfache Zufallsauswahlen werden auch innerhalb von Haushalten zur Auswahl von Personen anhand von Zufallszahlentabellen genutzt. Sie gewährleisten, dass jedes Haushaltsmitglied die gleiche Chance hat, an der Befragung teilzunehmen. Dazu dient eine Tabelle (*Schwedenschlüssel* oder *kish selection grid*), die für jede Haushaltsgröße eine zuvor ausgeloste Zufallszahl enthält. Bei Einpersonenhaushalten kann naturgemäß auch nur eine Person befragt werden, weshalb hier in der Tabelle immer eine 1 verzeichnet sein muss. Für Zweipersonenhaushalte wird eine 1 oder eine 2, für Dreipersonenhaushalte eine Zahl zwischen 1 und 3, für Vierpersonenhaushalte eine Zahl zwischen 1 und 4 usw. zufällig ausgewählt. Die Haushaltsgröße entspricht der Zahl der Personen, die zur Grundgesamtheit zählen. Als Auswahlkriterium wird das Alter herangezogen. Besteht ein Haushalt aus vier Personen und findet sich in der Tabelle für diese Haushaltsgröße die Zufallszahl „3“, so ist die drittälteste Person zu befragen. Die Intervieweranweisung könnte natürlich auch vorsehen, dass es die drittjüngste Person ist. Für jeden Haushalt wird ein eigener Schwedenschlüssel ausgelost. Die Problematik des Verfahrens liegt auf der Hand: Trifft der Interviewer die zu befragende Person nicht an, dann besteht die Gefahr, dass einfach eine anwesende Person ausgewählt wird.

9.3.2 Systematische Zufallsauswahlen

Da reine Zufallsauswahlen recht aufwändig sind, bedient man sich häufig systematischer Auswahlverfahren. Eine in der Praxis verbreitete Form der systematischen Zufallsauswahl ist der Zufallsweg (*random walk*), der im Zusammenhang mit der mehrstufigen Zufallsstichprobe erläutert wird.

Bei einer *systematischen Zufallsstichprobe* wird lediglich das *erste* Stichprobenelement per Zufall ermittelt (wiederum anhand der Zufallszahlentabelle oder des Zufallszahlengenerators). Ausgehend von dieser Zufallszahl werden alle weiteren Elemente systematisch ausgewählt. Dies geschieht, indem jedes k-te Element in die Stichprobe gelangt. Wie groß das Intervall k zwischen zwei auszuwählenden Elementen ist, hängt von der Größe der zu ziehenden Stichprobe und der Größe der Auswahlgesamtheit ab:

$$\text{Stichprobenintervall} \quad k = \frac{\text{Größe der Auswahlgesamtheit}}{\text{Größe der Stichprobe}}\,.$$

Ziel sei wiederum, eine Stichprobe mit 500 Schülern zu konstruieren. In unserem Fall würde sich nach dieser Formel ein Stichprobenintervall von $9700/500 = 19{,}4$ ergeben. Da wir nur jede 19. oder jede 20. Person auswählen können, runden wir ab und befragen ausgehend von der Zufallszahl jeden 19. Schüler. Der Bereich der Zufallszahl ergibt sich ebenfalls durch das Stichprobenintervall. Im Beispiel müssen wir eine Zahl zwischen 1 und 19 ermitteln. Wäre die Zufallszahl 18, dann ist die 18. Person auf der Liste die erste ausgewählte Person, danach folgt die 37., die 56. usw. bis zur 9689. Person der Liste. Insgesamt sind es 510 Personen (zehn mehr als beabsichtigt, da wir das Intervall abgerundet haben).

Von allen $\binom{N}{n}$ möglichen Stichproben des Umfangs n aus einer Grundgesamtheit N kann bei einer systematischen Zufallsauswahl nur eine geringe Zahl realisiert werden: sie entspricht dem Stichprobenintervall k. Im Beispiel könnten 19 verschiedene Stichproben gezogen werden, weil lediglich 19 verschiedene Zahlen (die Startzahlen) zufällig ausgewählt werden, während die „restlichen“ Stichprobenelemente in Abhängigkeit vom ersten ausgewählten Element bestimmt werden. Die einzelnen Ziehungen sind statistisch voneinander abhängig. Das erste gezogene Element bestimmt die Auswahl aller weiteren Elemente. Jede der k Stichproben (nicht aber

jede der *möglichen* Stichproben) und jedes Element der Auswahlgesamtheit besitzt die gleiche Wahrscheinlichkeit, ausgewählt zu werden (vgl. zum Stichprobenfehler Levy und Lemeshow 1991, 74).

Problematisch ist eine systematische Zufallsauswahl, wenn die Systematik der Auswahl mit der Ordnung der Kartei übereinstimmt und das Ordnungskriterium mit dem Untersuchungsmerkmal korreliert (Böltken 1976, 166-173). Wäre (zugegebenermaßen unrealistisch) unsere Kartei so aufgebaut, dass der zufällig ausgewählte erste Schüler Mathematik als Leistungsfach hätte und jeder weitere 19. Schüler ebenfalls, dann enthielte unsere Auswahl lediglich Schüler mit mathematischem Schwerpunkt. Als Folge erhielten wir ein falsches Bild der Studienwünsche, weil Schüler mit Leistungskurs Mathematik sich in ihrer Neigung zu bestimmten Studienfächern mit hoher Wahrscheinlichkeit von allen Schülern unterscheiden.

9.3.3 Komplexe Zufallsauswahlen

Liegt keine Auflistung aller Einheiten der Auswahlgesamtheit vor – wie im Falle der 60 Millionen bundesdeutschen Wahlberechtigten –, verwendet man zweckmäßigerweise ein *mehrstufiges Stichprobenverfahren.* Soll eine selten vorkommende Merkmalsausprägung untersucht werden, dann bietet sich eine *disproportional geschichtete Stichprobe* an, in der Personen mit der interessierenden Merkmalsausprägung überproportional vertreten sind. Ist der räumliche oder personelle Kontext einer Zielperson für eine Untersuchung interessant, dann ist eine *Klumpenstichprobe* angemessen.

Geschichtete Auswahlen

Zur Ziehung einer geschichteten Stichprobe werden die Elemente der Auswahlgesamtheit bezüglich des interessierenden Merkmals in Schichten (bzw. Gruppen) eingeteilt. Aus diesen Schichten werden dann (getrennt) Zufallsstichproben gezogen.

Für eine geschichtete Stichprobe sprechen zwei Gründe: Besteht die Auswahlgesamtheit aus verschiedenen Gruppen, die in sich sehr homogen sind, dann kann die Genauigkeit der Stichprobe gegenüber einer einfachen Zufallsstichprobe bei gleicher Stichprobengröße erhöht werden. Wird die Größe einer Schicht entsprechend ihrem Anteil an der Grundgesamtheit gewählt, dann spricht man von einer *proportional geschichteten Stichprobe.*

In der Regel entscheidet man sich jedoch für geschichtete Stichproben, wenn eine oder mehrere Ausprägungen des Schichtmerkmals bei einer einfachen Zufallsauswahl nicht in hinreichender Zahl in der Stichprobe vertreten wären. In solchen Fällen zieht man eine *disproportional geschichtete Stichprobe*, in der die Anteile der einzelnen Schichten *nicht* den Anteilen in der Grundgesamtheit entsprechen. Die interessierende Schicht wird überrepräsentiert. Im Gegensatz zu den zuvor besprochenen einfachen Zufallsauswahlen hat hier jedes Element nicht mehr die gleiche, sondern nur noch eine *bekannte* bzw. *angebbare* Chance, in die Stichprobe zu gelangen. Im ALLBUS wird für West- und Ostdeutschland die Stichprobenziehung getrennt vorgenommen, und zwar so, dass die Bevölkerung Ostdeutschlands in der Gesamtstichprobe im Vergleich zur Grundgesamtheit überrepräsentiert ist. Durch die disproportionale Schichtung wird erreicht, dass die Fallzahlen für separate Analysen der ostdeutschen Befragten ausreichend hoch sind. Wertet man beide Stichproben zusammen aus, dann muss die unterschiedliche Auswahlwahrscheinlichkeit für Ost- und Westdeutsche wieder rückgängig gemacht werden. Dies geschieht durch eine Design-Gewichtung (vgl. für den ALLBUS 1996 Wasmer et al. 1996, 61 f.). Beim ALLBUS wird innerhalb der Schichten – in Ost- und Westdeutschland – eine mehrstufige Zufallsstichprobe gezogen.

Geschichtete Zufallsauswahlen setzen voraus, dass die Verteilung des Schichtmerkmals in der Grundgesamtheit bekannt ist. Außerdem muss für jede Auswahleinheit das Schichtungsmerkmal feststellbar sein.

Klumpenauswahlen

Klumpenstichproben bieten sich immer dann an, wenn man den Kontext, also Gruppenzusammenhänge, untersuchen möchte. Die Auswahl bezieht sich nicht auf Untersuchungseinheiten, sondern auf *Aggregate von Untersuchungseinheiten*, so genannte Klumpen. Von einer Klumpenauswahl spricht man nur dann, wenn *alle* Elemente eines Klumpens in die Stichprobe gelangen und die Elemente des Klumpens die Untersuchungseinheiten sind. Wenn wir die Vermutung haben, dass die Studienwünsche der einzelnen Abiturienten von den Studienwünschen ihrer Mitschüler abhängen, wäre eine Klumpenstichprobe angemessen. Die Klumpen wären in diesem Fall Schulklassen. Es würde also eine Auswahl von Schulklassen getroffen. Alle Schüler der ausgewählten Klassen wären in der Stichprobe.

Klumpenauswahlen haben bei gleicher Stichprobengröße in der Regel einen größeren Stichprobenfehler als einfache Zufallsauswahlen. Vor allem dann, wenn die Klumpen in sich sehr homogen sind, sich aber stark voneinander unterscheiden. Wären die Gymnasien beispielsweise stark fachlich ausgerichtet (technische Gymnasien, Wirtschaftsgymnasien, Gymnasium mit fremdsprachigem Schwerpunkt), dann dürften sich die Studienwünsche von Abiturienten von Gymnasien mit verschiedenen Schwerpunkten stark unterscheiden. Würde z. B. eine Klasse eines Wirtschaftsgymnasiums ausgewählt, dann könnte man annehmen, dass sich die Studienwünsche der Schüler erheblich zugunsten wirtschaftswissenschaftlicher Studienzweige von allen Schülern unterscheiden. Wir hätten also eventuell einen für die Grundgesamtheit ‚untypischen' Klumpen gezogen. Da alle Schüler dieser Klasse in die Stichprobe gelangen, fällt die Abweichung erheblich ins Gewicht.

Mehrstufige Zufallsauswahlen

Mehrstufige Auswahlen sind eine Reihe *nacheinander durchgeführter Zufallsauswahlen*. Auf der ersten Stufe wird eine Stichprobe aus *Gruppen von Elementen*, den *Primäreinheiten*, gezogen. Primäreinheiten sind häufig regionale Einheiten wie Gemeinden oder Stimmbezirke. Aus den Elementen der ausgewählten Primäreinheiten wird dann eine weitere Stichprobe gezogen. Diese Elemente sind die *Sekundäreinheiten*. Die ausgewählten Sekundäreinheiten können Grundlage einer weiteren Stichprobenziehung sein. Auf der letzten Auswahlstufe werden die Untersuchungseinheiten ausgewählt.

Das Vorgehen soll zunächst am Beispiel der Studienwünsche rheinland-pfälzischer Schüler verdeutlicht werden: Wir könnten zunächst eine Stichprobe aus allen Schulen ziehen, die ein 13. Schuljahr anbieten (Gymnasien, integrierte Gesamtschulen). Die Auswahlgesamtheit besteht auf der ersten Stufe aus den 146 Schulen, an denen das Abitur erworben werden kann (Primäreinheiten). Aus den ausgewählten Schulen werden dann auf der zweiten Stufe Schüler (Sekundäreinheiten) ausgewählt. Wie viele Schüler wir pro Schule auswählen müssen, um eine Stichprobengröße von 500 Schülern zu erreichen, hängt davon ab, wie viele Schulen auf der ersten Stufe ausgewählt wurden. Wenn wir 50 Schulen auswählen, müssten pro Schule 10 Schüler ausgewählt werden; wählen wir 25 Schulen aus, dann erhöht sich die Zahl der Schüler auf 20.

Der Nachteil mehrstufiger Auswahlverfahren besteht darin, dass man auf jeder Auswahlstufe einen Stichprobenfehler begeht und sich diese Stichprobenfehler addieren. Der Stichprobenfehler wird umso kleiner, je größer die Stichprobe und je geringer die Varianz des interessierenden Merkmals in der Auswahlgesamtheit ist. Um den Stichprobenfehler auf der ersten Auswahlstufe klein zu halten, müssten also möglichst viele Gruppen bzw. Primäreinheiten ausgewählt werden. Ebenso kann man den Stichprobenfehler der zweiten Stufe minimieren, indem möglichst viele Sekundäreinheiten ausgewählt werden. Im Beispiel müssten zunächst möglichst viele Schulen ausgewählt werden und aus diesen Schulen wiederum möglichst viele Schüler. In diesem Fall minimiert man beide Fehler durch die Vergrößerung der Stichprobe.

Bei gegebener Stichprobengröße, die bei uns 500 betragen soll, ist es jedoch unmöglich, gleichzeitig beide Stichprobenfehler zu minimieren: Je mehr Schulen ausgewählt werden, umso weniger Schüler müssen pro Schule befragt werden und umgekehrt. Anders ausgedrückt: Indem man den Stichprobenfehler einer Stufe reduziert, erhöht man den Stichprobenfehler auf einer anderen Stufe.

Aus dieser Zwickmühle kann man sich jedoch befreien, wenn man zusätzlich die Homogenität der Auswahlgesamtheit berücksichtigt. Die Schüler einer Schule sind einander ähnlicher als die Schüler verschiedener Schulen. Wählen wir nur wenige Schulen aus, besteht eine größere Gefahr, dass die ausgewählten Schüler untypisch für alle Schüler sind, als wenn wir möglichst viele Schulen in der Stichprobe berücksichtigen, diese jedoch immer nur durch wenige Schüler repräsentiert werden. Eine möglichst hohe Zahl an auszuwählenden Primäreinheiten schmälert allerdings den Effizienzvorteil mehrstufiger Stichproben, da wir dann wiederum mehr Schulen um Schülerlisten bitten müssen, die Interviewer (wenn wir die Befragung mündlich vornehmen) weitere Wege zurücklegen müssen usw. Man wägt in der Regel die Effizienzvorteile weniger Primäreinheiten gegen die Nachteile einer ungenaueren Stichprobe ab und wird einen Mittelweg beschreiten.

Da die *Primäreinheiten* – im Beispiel Schulen – in der Regel unterschiedlich groß sind, müssen diese mit einer Wahrscheinlichkeit ausgewählt werden, die *proportional zur ihrer Größe* ist. Man bezeichnet dieses Design auch als **PPS-Design** (probability proportional to size).

Nehmen wir an, wir haben uns entschieden, 500 Abiturienten auszuwählen. Auf der ersten Stufe sollen 25 Schulen ausgewählt werden, auf der zweiten

Stufe jeweils 20 Abiturienten. Die Berechnung der Auswahlwahrscheinlichkeit soll nun für zwei verschiedene Schulen verdeutlicht werden: Schule A hat 189 Abiturienten, Schule B 49. Würden wir auf der ersten Stufe die Schulen nicht entsprechend ihrer Größe auswählen, so hätten beide Schulen eine Auswahlwahrscheinlichkeit von $25 \times 1/146 = 25/146 = 0{,}171$, da es 146 Schulen gibt und wir 25 Schulen auswählen. Auf der zweiten Stufe hätte ein Abiturient innerhalb der Schule A die Chance $20/189 = 0{,}106$ ausgewählt zu werden, innerhalb der Schule B wäre die Chance für einen Abiturienten $20/49 = 0{,}408$. Die Gesamtwahrscheinlichkeit kann man berechnen, indem die Wahrscheinlichkeiten beider Stufen multipliziert werden. Die Chance, dass ein Abiturient der Schule A in die Stichprobe gelangt, würde insgesamt also 0,018 ($0{,}171 \times 0{,}106$) betragen, während ein Abiturient der Schule B mit einer Wahrscheinlichkeit von 0,07 ($0{,}171 \times 0{,}408$) in der Auswahl vertreten sein würde.

Wie man sieht, resultieren die ungleichen Auswahlwahrscheinlichkeiten der Schüler aus der unterschiedlichen Zahl der Abiturienten an den beiden Schulen. Um die unterschiedlichen Auswahlwahrscheinlichkeiten auf der zweiten Stufe auszugleichen, muss eine Schule mit 189 Abiturienten eine größere Wahrscheinlichkeit erhalten, in die Stichprobe zu gelangen, als eine Schule mit 49 Abiturienten. Die Auswahl einer Schule muss proportional zu ihrer Größe erfolgen, wobei die Größe einer Schule ihrem Anteil an allen Abiturienten entspricht:

$$\text{Größe} = \frac{\text{Zahl der Abiturienten einer Schule}}{\text{Zahl aller Abiturienten in Rheinland-Pfalz}}\,.$$

Für Schule A bedeutet dies, dass ihre Auswahlwahrscheinlichkeit von $25 \times 1/146 = 25/146 = 0{,}171$ auf $25 \times 189/9700 = \mathbf{0{,}487}$ *steigt*, für Schule B, dass ihre Auswahlwahrscheinlichkeit von $25 \times 1/146 = 25/146 = 0{,}171$ auf $25 \times 49/9700 = \mathbf{0{,}126}$ *sinkt* (vgl. Tabelle 9.5).

Die Wahrscheinlichkeit, dass ein Abiturient der Schule A in die Stichprobe gelangt, beträgt jetzt 0,052 ($0{,}487 \times 0{,}106$); die eines Abiturienten der Schule B beträgt ebenfalls 0,052 ($0{,}126 \times 0{,}408$).[1] Die Wahrscheinlichkeit in die Stichprobe zu gelangen, ist für (die kleinere) Schule B zwar viel

1 Abweichungen können aufgrund von Rundungen entstehen. Beim Nachrechnen bitte ungerundete Werte verwenden.

geringer als für (die größere) Schule A (0,126 zu 0,487); dies wird aber durch die höhere Auswahlwahrscheinlichkeit der Schüler von Schule B auf der zweiten Stufe kompensiert.

Tabelle 9.5: Auswahlwahrscheinlichkeit beim PPS-Design

	Schule A	Schule B
1. Stufe	$25 \times 189/9700 = \mathbf{0{,}487}$	$25 \times 49/9700 = \mathbf{0{,}126}$
2. Stufe	$20/189 = \mathbf{0{,}106}$	$20/49 = \mathbf{0{,}408}$
Insgesamt	$0{,}487 \times 0{,}106 = \mathbf{0{,}052}$	$0{,}126 \times 0{,}408 = \mathbf{0{,}052}$

Für die Bundesrepublik existiert kein zentrales Einwohnermelderegister, das die Ziehung einer einfachen Zufallsstichprobe ermöglicht. Für Bevölkerungsumfragen setzt man daher mehrstufige Auswahlverfahren ein.

Eine Möglichkeit ist eine Auswahl auf Basis des ADM-Mastersamples. ADM steht für Arbeitskreis Deutscher Marktforschungsinstitute (vgl. Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute 1999). Für das ADM-Mastersample wurde aus den Stimmbezirken zur Wahl des Deutschen Bundestags eine Stichprobe gezogen (vgl. Porst 1985, 86–88). Aus diesem Mastersample wurden Unterstichproben – Netze – gezogen, die jeweils 210 Stimmbezirke (*Sample-Points*) umfassen. Die Stimmbezirke stellen die Primäreinheiten dar. Auf der zweiten Stufe verzichtet man auf die Erstellung einer Liste mit allen Haushalten. Die Zufallsauswahl der Haushalte soll durch das Beschreiten eines *Zufallswegs* (*Random Route*) gewährleistet werden. Der Interviewer erhält eine zufällig gezogene Startadresse und bestimmte Begehungsregeln, die zur Ermittlung der weiteren Adressen führen. Adressenermittlung und Interview können dabei getrennt (*Adress Random*) oder gemeinsam erfolgen. Auf der dritten Stufe werden aus den Haushalten die Befragungspersonen ausgewählt. Besteht ein Haushalt aus mehreren Personen, dann muss die zu interviewende Person ermittelt werden, z. B. die Person, die zuletzt Geburtstag hatte. Häufig werden aber auch Zufallszahlentabellen wie der Schwedenschlüssel verwendet. Die Auswahlwahrscheinlichkeiten auf der letzten Stufe sind ungleich: eine Person in einem Einpersonenhaushalt hat eine Auswahlwahrscheinlichkeit von 1, für Personen in Vierpersonenhaushalten beträgt diese 1/4. Die unterschiedlichen Auswahlwahrscheinlichkeiten können durch Gewichtung bei der Datenanalyse wieder rückgängig gemacht werden (Design-

Gewichtung). Der Nachteil der Auswahl der Haushalte und der Befragungspersonen durch die Interviewer liegt auf der Hand. Eine Zufallsauswahl liegt nur dann vor, wenn der Interviewer sich sowohl an die Begehungsregeln als auch an die Zufallszahlentabelle zur Auswahl des Befragten hält. Hier ist die Gefahr groß, dass die Interviewer aus Zeitgründen die Begehungsregeln „abkürzen", oder die Person befragen, die sie gerade im Haushalt antreffen (vgl. Dorroch 1994).

Das ADM-Mastersample wurde auch zur Ziehung der Stichproben des Sozio-ökonomischen Panels herangezogen. Der 1984 gezogenen Stichprobe A des Sozio-ökonomischen Panels liegen z. B. knapp 600 Sample-Points zugrunde (vgl. Hanefeld 1987, 171–175, 181 f.). Beim SOEP erfolgte die Ermittlung der Haushaltsadressen getrennt vom Interview (Adress Random). Da das SOEP eine Haushaltsstichprobe ist, entfiel die dritte Stufe (vgl. Hanefeld 1987, 136). Bis 1992 (außerdem 1998) wurde das ADM-Stichprobendesign auch für den ALLBUS verwendet.

Seit 1994 (Ausnahme: 1998) werden bei den ALLBUS-Erhebungen Gemeindestichproben mit Adressenziehung aus den Einwohnermelderegistern realisiert. Die Stichprobenziehung erfolgt getrennt für Ost- und Westdeutschland (disproportionale Schichtung). Auf der ersten Stufe werden proportional zur Bevölkerungsgröße Gemeinden ausgewählt. Aus den Einwohnermelderegistern dieser Gemeinden werden auf der zweiten Stufe zufällig Personen gezogen (vgl. zum ALLBUS 1994 Koch et al. 1994). Die Befragung erfolgt persönlich-mündlich. Der Vorteil gegenüber dem ADM-Design liegt unter anderem darin, dass die Stichprobenziehung vollkommen getrennt von der Feldphase ist. Die Interviewer haben keinen Einfluss auf die Auswahl der Personen. Einwohnermelderegisterstichproben sind allerdings teuer: auf eine Registerstichprobe musste beim ALLBUS 1998 aus finanziellen Gründen verzichtet werden (vgl. Koch et al. 1999, 2 f.).

Telefonstichproben

Auch bei Telefonumfragen wird häufig eine mehrstufige Auswahl realisiert. Die Zufallsauswahl kann ohne eine Liste der Telefonnummern erfolgen. Das ursprünglich in den USA entwickelte Verfahren heißt *Random Digit Dialling* (RDD). Auf der ersten Stufe kann eine Region/Vermittlungsstelle ausgewählt werden, auf der nächsten Stufe wird per Zufallsverfahren eine Anschlussnummer erzeugt. Die zufällige Generierung von Telefonnummern hat gegenüber einer Auswahl aus dem Telefonbuch den Vorteil, dass auch

Teilnehmer in die Stichprobe gelangen können, die nicht im Telefonbuch verzeichnet sind. Dies ist in den USA von besonderer Bedeutung, da dort der Anteil nicht eingetragener Telefonnummern sehr hoch ist. In Deutschland sind nach Angaben von Follmer und Smid (1998, 49) 10,6 % der west- und 18,7 % der ostdeutschen Anschlüsse nicht im Telefonbuch gelistet.

Die Anwendung des RDD scheitert in der Bundesrepublik daran, dass die Vorwahlen und Teilnehmernummern hier – im Gegensatz zu den USA – unterschiedliche Längen haben. In Deutschland hat sich für wissenschaftliche Untersuchungen der Gabler-Häder-Auswahlrahmen (Gabler und Häder 1998) und in der Marktforschung das ADM-Telefonmastersample durchgesetzt (Häder und Glemser 2006). Beim Verfahren von Gabler und Häder werden zunächst die 100er Blocks in Ortsnetzbereichen ermittelt, in denen sich mindestens eine eingetragene Telefonnummer befindet. Ein 100er Block ist der Stamm einer Telefonnummer ohne die letzten beiden Ziffern. Es wird davon ausgegangen, dass in den 100er Blocks, in denen sich keine eingetragenen Telefonnummern befinden, auch keine nicht eingetragenen Nummern existieren. Für die besetzten 100er Blocks werden alle möglichen Ziffernfolgen erzeugt. Diese Ziffernfolgen werden als Auswahlgesamtheit für Telefonstichproben zur Verfügung gestellt. Eingetragene und nicht eingetragene Telefonnummern haben beim Gabler-Häder-Design die gleiche Auswahlwahrscheinlichkeit. Nicht jede der generierten Ziffern repräsentiert eine existierende Telefonnummer. Es kann sich auch um eine nicht vergebene Nummer handeln. Die nicht vergebenen Nummern sind *stichprobenneutrale* Ausfälle. Telefonstichproben stellen Haushaltsstichproben dar. Kommt ein Kontakt zustande, dann muss auch hier wieder zufällig eine Person ausgewählt werden.

Telefonstichproben für Bevölkerungsumfragen setzen eine hinreichend hohe Telefondichte voraus. Dies ist in West- und auch in Ostdeutschland der Fall. Problematisch ist allerdings der zunehmende Anteil von Privathaushalten, die ausschließlich über ein Mobiltelefon erreichbar sind. Bei einer Auswahl aus den eingetragenen und nicht eingetragenen Festnetznummern ist deren Auswahlwahrscheinlichkeit null, d. h. sie können nicht in die Stichprobe gelangen.

In diesem Kapitel wurden die Prinzipien der verschiedenen Auswahlverfahren dargestellt. In der Praxis ist die Ziehung einer Zufallsstichprobe aufwändig und mit zahlreichen Problemen behaftet. Einen Einblick in die

Praxis der Stichprobenziehung vermitteln Gabler et al. (1998) und der Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute (1999).

9.4 Nicht zufallsgesteuerte Auswahlverfahren

Bei nicht zufallsgesteuerten Auswahlverfahren unterliegt die Auswahl der Untersuchungseinheiten keinem Zufallsprozess. Unterschieden wird zwischen *willkürlichen* und *bewussten* Auswahlverfahren. Willkürlich ist ein Verfahren, wenn keine Auswahlkriterien angegeben werden. Bei bewussten Auswahlen werden die Elemente nach bestimmten Zielen ausgewählt.

Im Gegensatz zu den zufallsgesteuerten Auswahlverfahren ist der statistische Schluss auf eine Grundgesamtheit nicht möglich, weil keine Angabe über die Auswahlwahrscheinlichkeit gemacht werden kann. Wissenschaftler sind jedoch nicht immer an Aussagen über die Grundgesamtheit interessiert. Dies gilt insbesondere für die qualitative Sozialforschung (vgl. Flick 2007). Zudem sind zufällige Auswahlverfahren nicht immer realisierbar. Die Ziehung einer Zufallsstichprobe aus den wohnungslosen Menschen in der Bundesrepublik dürfte sich schwierig gestalten. Möglich sind aber Erhebungen in Einrichtungen für wohnungslose Menschen. Die Aussagen der Untersuchung müssen sich dann auf die untersuchte Gruppe beschränken. Auch experimentelle Designs benötigen keine Zufallsstichproben. In Experimenten ist die zufällige Zuteilung der Teilnehmer auf die einzelnen Versuchsgruppen entscheidend, nicht eine Zufallsauswahl der Teilnehmer aus einer Grundgesamtheit.

Auch bei nicht zufallsgesteuerten Auswahlen müssen Kriterien der Auswahl angebbar und beurteilbar sein. Willkürliche Auswahlen – von Schnell et al. (2008) als Auswahlen aufs Geratewohl charakterisiert – sind wissenschaftlich nicht zu rechtfertigen. Zu den willkürlichen Auswahlen zählen Passantenbefragungen. Eine willkürliche Auswahl liegt aber auch dann vor, wenn im Rahmen einer qualitativen Erhebung ausschließlich die leicht verfügbaren Fälle (und beispielsweise nicht die theoretisch interessanten Fälle) untersucht werden.

Quotenauswahl

Eine Mischform aus bewusster und willkürlicher Auswahl stellen Quotenauswahlen dar. In wissenschaftlichen Untersuchungen werden diese vor

allem vom Institut für Demoskopie (IfD) in Allensbach verwandt. Mit Hilfe der Quotenauswahl wird eine Stichprobe angestrebt, die Aussagen über die Grundgesamtheit ermöglicht.

Dies soll durch die Vorgabe von *Quoten*, d. h. *Anteile, mit denen bestimmte Merkmalsausprägungen in der Stichprobe vorhanden sein sollen*, erreicht werden. Die Anteile dieser Merkmalsausprägungen müssen in der Stichprobe genauso groß sein wie in der Grundgesamtheit. Um Quoten vorgeben zu können, muss man natürlich zunächst wissen, wie groß der Anteil eines Merkmals in der Grundgesamtheit ist. Anhand dieser Quoten wählt der Interviewer dann *willkürlich* die Befragten aus. Die „Willkür“ des Interviewers ist nicht mit *Zufall* gleichzusetzen, wie von den Befürwortern dieses Auswahlverfahrens behauptet wird (vgl. Noelle-Neumann und Petersen 1996, 259).

Kommen wir noch einmal auf die Umfrage unter rheinland-pfälzischen Abiturienten zurück. Als Quotierungsmerkmal könnte man z. B. das Geschlecht vorgeben. Wären 55 % aller rheinland-pfälzischen Abiturienten Männer und 45 % Frauen, dann müßten in unserer Stichprobe ebenfalls 55 % Männer und 45 % Frauen vertreten sein. Welche Schüler wir befragen, ist beliebig, solange wir uns an die vorgegebenen Quoten halten. Um dem angestrebten Ideal einer „repräsentativen“ Stichprobe näher zu kommen, könnte man zusätzlich zum Geschlecht noch vorgeben, wie viel Prozent der Schüler aus verschiedenen sozialen Schichten kommen.

Das Geschlecht und die Schichtzugehörigkeit können nun als unabhängige Quoten, d. h. isolierte Merkmalsausprägungen, vorgegeben werden. Z. B. könnte die Quote lauten „60 % Frauen und 40 % Männer sowie 30 % Arbeiterschicht, 60 % Mittelschicht und 10 % Oberschicht“. Wenn die Quotierung unabhängig voneinander erfolgt, kann es bei diesem Beispiel theoretisch passieren, dass alle ausgewählten Mittelschichtangehörigen auch Frauen sind und die Männer alle Arbeiter- und Oberschichtangehörige sind. Die einzelnen Quoten für Geschlecht und Schichtzugehörigkeit wären damit zwar erfüllt, nicht aber eine *kombinierte* bzw. *abhängige* Quote, die die gemeinsame Verteilung des Geschlechts und der Schichtzugehörigkeit berücksichtigt. Abhängige Quoten setzen exakte Kenntnisse der Grundgesamtheit voraus. Man kann sich zudem leicht vorstellen, wie unrealisierbar abhängige Quoten werden, wenn mehr als zwei Merkmale bei der Quotierung kombiniert werden. Zudem ist die Gefahr des „Umdefinierens“ durch Interviewer bei seltenen Merkmalskombinationen groß, weil

der Aufwand für den Interviewer steigt (vgl. Dorroch 1994, 40). Aus diesem Grunde werden in der Regel meist unabhängige Quoten verwendet bzw. kombinierte Quoten für zwei Merkmale (vgl. Noelle-Neumann und Petersen 1996, 257).

Mit der Quotenauswahl sind eine Reihe von Nachteilen verbunden:

- Die Verteilung der Quotierungsmerkmale in der Grundgesamtheit muss bekannt sein. Dies ist nur für wenige Merkmalsausprägungen der Fall.
- Die Quotierungsmerkmale müssen für den Interviewer leicht erfassbar sein, wie etwa das Geschlecht. Schon das Alter, die Schichtzugehörigkeit oder die Schulbildung können nicht mehr per Augenschein festgestellt werden.
- Ausfälle werden verdeckt, da der Interviewer bei Befragungsverweigerungen einfach die nächste Person mit den geforderten Merkmalen sucht. Eine Statistik über Ausfälle sowie die Berechnung der Ausschöpfungsquote ist somit nicht möglich, die daraus resultierenden Fehler bleiben unbekannt. Die Argumentation, Quotenauswahlen seien nicht schlechter als Zufallsstichproben, weil dort hohe Anteile von Ausfällen zu verzeichnen seien (vgl. Noelle-Neumann und Petersen 1996, 267 f.), ist deshalb nicht stichhaltig. Befragungsverweigerungen treten auch bei Quotenauswahlen auf. Sie werden nur nicht dokumentiert.
- Quotenstichproben sind möglicherweise ‚repräsentativ' im Hinblick auf die quotierten Merkmale. Ob die Stichprobe auch bei anderen Merkmalen ein verkleinertes Abbild der Grundgesamtheit darstellt, ist nicht bekannt. Bei Zufallsauswahlen sorgt der Zufallsprozess dafür, dass seltene Merkmale nur eine geringe Wahrscheinlichkeit haben in die Stichprobe zu gelangen und häufige Merkmale eine hohe Wahrscheinlichkeit. Bei Quotenauswahlen ist dies nicht der Fall. Es sind keine Auswahlwahrscheinlichkeiten angebbar. Der Schluss auf die Grundgesamtheit ist statistisch daher nicht begründbar.

Der Grund für den Einsatz von Quotenauswahlen besteht vor allem darin, dass diese kostengünstiger als Zufallsauswahlen sind.

Aufgaben zu Auswahlverfahren

1. Warum benötigt man Stichproben? Schildern Sie deren Vor- und Nachteile im Vergleich zu Vollerhebungen.
2. Für den ALLBUS 1994 wurden zunächst 151 Gemeinden ausgewählt. Aus den Einwohnermelderegistern dieser Gemeinden wurden die Adressen der zu befragenden Personen per Zufallsauswahl ermittelt. Welches Stichprobenverfahren wurde angewandt?
3. Sie möchten wissen, in welchem Umfang das „Studi-Ticket" von den Mainzer Studierenden genutzt wird. Das Studentensekretariat stellt Ihnen dazu eine Liste mit den Namen der 28.734 Studierenden (WS 1995/1996) zur Verfügung. Ihre Stichprobe soll mindestens 1.000 Studierende umfassen.
   Was ist in diesem Fall die Grundgesamtheit, was die Auswahlgesamtheit? Geben Sie Beispiele für undercoverage und overcoverage.
   Aus dieser Liste möchten Sie nun eine systematische Zufallsstichprobe ziehen. Wie gehen Sie vor?
4. Worin besteht der Unterschied zwischen zufallsgesteuerten und nicht zufallsgesteuerten Auswahlverfahren?

10 Wahrscheinlichkeitsverteilungen

Mit Hilfe von Wahrscheinlichkeitsverteilungen lässt sich angeben, wie Stichprobenkennwerte vom wahren Wert der Grundgesamtheit abweichen. Diese Überlegungen sind notwendig, um die Frage zu beantworten, wie auf Basis einer einzigen Stichprobe auf die Grundgesamtheit geschlossen (Kapitel 11) und Hypothesen über die Grundgesamtheit getestet werden können (Kapitel 12).

Wie im Kapitel Stichprobenziehung angedeutet wurde, muss das Stichprobendesign bei der Datenanalyse berücksichtigt werden: Design-Gewichte werden notwendig, wenn nicht jedes Element der Grundgesamtheit die gleiche Auswahlwahrscheinlichkeit hatte – z. B. bei disproportional geschichteten Stichproben. Bei Schätzung des Stichprobenfehlers muss eine Systematik der Auswahl, eine Schichtung oder die „Klumpung" von Fällen berücksichtigt werden (Kohler 2006). Mit Statistik-Programmen wie Stata ist dies einfach möglich. Eine Einführung in die Analyse systematischer und komplexer Zufallsstichproben geben Scheaffer et al. (1996) (siehe auch Levy und Lemeshow 1991). Zur Vereinfachung wird in den folgenden Kapiteln von *einfachen* Zufallsstichproben ausgegangen.

10.1 Relative Häufigkeit und Wahrscheinlichkeit

In Kapitel 9.2 haben wir verschiedene Zufallsexperimente betrachtet: das Werfen eines Würfels und das Ziehen einfacher Zufallsstichproben. Bei gleichwahrscheinlichen Ereignissen können wir die Wahrscheinlichkeit des Auftretens von Ereignissen theoretisch bestimmen, wie wir gesehen haben (*Laplace-Wahrscheinlichkeit* bzw. klassische Wahrscheinlichkeit). So beträgt die Wahrscheinlichkeit, beim Würfeln eine 6 zu erzielen, genau 1/6.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_10,

Wahrscheinlichkeiten sind mit relativen Häufigkeiten eng verknüpft, was im **Bernoulli-Theorem** ausgedrückt wird. Das Bernoulli-Theorem besagt, dass die relative Häufigkeit eines Ereignisses bei unendlicher Wiederholung des Zufallsexperimentes der Wahrscheinlichkeit entspricht. Man kann davon ausgehen, dass anstelle einer „unendlich häufigen" Wiederholung auch eine „sehr häufige" Wiederholung eines Zufallsvorgangs genügt, also z. B. 100-mal oder 1.000-mal. Das Bernoulli-Theorem ist eine Anwendung des Gesetzes großer Zahlen (vgl. Fahrmeir et al. 2007, 313 f.).

Dem Bernoulli-Theorem liegt ein statistischer bzw. frequentistischer Wahrscheinlichkeitsbegriff zugrunde (Gleichung 10.1). Beim 100-maligen Werfen eines Würfels sollte jede Augenzahl ungefähr 16 oder 17 Mal auftreten, denn die Wahrscheinlichkeit für jede der 6 Zahlen beträgt $1/6 = 0{,}1\bar{6}$ und $100 \times 0{,}1\bar{6} = 16{,}\bar{6}$. Bei 1.000 Würfen sollten demnach ca. 167 Würfe auf eine Augenzahl entfallen. Eine Voraussetzung des Bernoulli-Theorems ist, dass eine unendlich häufige Wiederholung theoretisch möglich ist.

$$P(A) = \lim_{n \to \infty} \text{rel. Häufigkeit}(A) \tag{10.1}$$

Das Bernoulli-Theorem lässt sich anhand eines Experiments illustrieren. Wir bezeichnen dieses Experiment als `Experiment I`, da wir später weitere Zufallsexperimente durchführen werden. Ein Programm, das Zufallszahlen erzeugt, ersetzt dabei den Würfel. Ein solcher „Zufallszahlengenerator" lässt sich so konstruieren, dass er eine beliebige Zahl innerhalb eines gegebenen Intervalls mit einer bestimmten Wahrscheinlichkeit produziert. Wir lassen uns eine der Zahlen 1 bis 6 erzeugen, wobei jede dieser Zahlen gleich wahrscheinlich ist. Das entspricht dem Werfen mit einem Würfel. Alle Simulationen in diesem Buch wurden mit den Programmen GSTAT und GSTAT2 von Fred Böker (1993, 1998) durchgeführt.

In Tabelle 10.1 werden die Ergebnisse dieses `Experiments I` zusammengefasst. In Spalte A sind die möglichen Ereignisse des Zufallsexperiments *Werfen eines Würfels* angegeben, die Augenzahlen 1, 2, 3, 4, 5, 6. In der Spalte $P(A)$ finden sich die Wahrscheinlichkeiten der Ereignisse. Diese lassen sich hier einfach durch Auszählen der günstigen im Vergleich zu allen Ereignissen bestimmen (*Laplace-Wahrscheinlichkeit*). Die Wahrscheinlichkeit ist also bekannt. In den restlichen Spalten der Tabelle sind

die relativen Häufigkeiten der einzelnen Augenzahlen angegeben, wenn der Würfel 10-mal, 50-mal, 100-mal usw. bis 1.000.000-mal geworfen wurde. Beispielsweise gibt die Zahl 0,3000 in der Spalte ‚10' die relative Häufigkeit für die Augenzahl 6 an. Die 6 hat also einen Anteil von 0,3 bzw. 30 % bei den 10 Würfen. Die absoluten Häufigkeiten erhält man, wenn man die relativen Häufigkeiten mit der jeweiligen Zahl der Würfe multipliziert. Die 6 ist bei den 10 Würfen also 3-mal aufgetreten. Wie man sieht, entsprechen die relativen Häufigkeiten bei weniger als 100 Würfen nur sehr ungenau den Wahrscheinlichkeiten $P(A)$, nähern sich diesen aber mit größer werdender Zahl von Würfen immer mehr an. Bereits bei 10.000 Würfen stimmen die relativen Häufigkeiten mit den Wahrscheinlichkeiten bis auf die zweite Nachkommastelle überein.

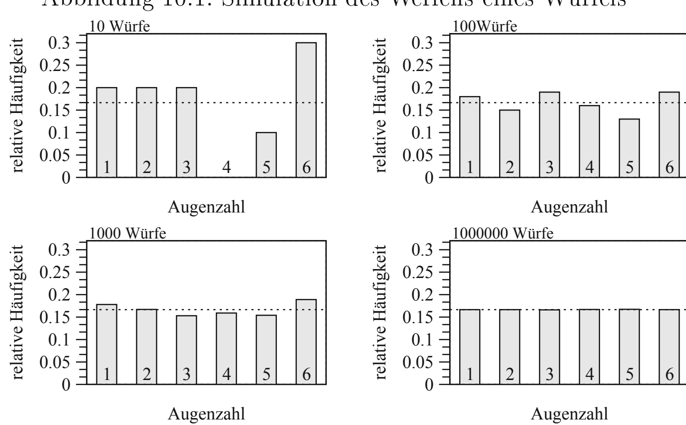
Tabelle 10.1: Wahrscheinlichkeit und relative Häufigkeit beim Werfen eines Würfels

		Anzahl der Würfe						
A	$P(A)$	10	50	100	1.000	10.000	100.000	1.000.000
1	$0{,}1\bar{6}$	0,2000	0,1400	0,1800	0,1780	0,1674	0,1668	0,1666
2	$0{,}1\bar{6}$	0,2000	0,1600	0,1500	0,1670	0,1676	0,1651	0,1666
3	$0{,}1\bar{6}$	0,2000	0,2200	0,1900	0,1530	0,1637	0,1673	0,1660
4	$0{,}1\bar{6}$	0,0000	0,1800	0,1600	0,1590	0,1680	0,1672	0,1669
5	$0{,}1\bar{6}$	0,1000	0,1000	0,1300	0,1540	0,1656	0,1683	0,1673
6	$0{,}1\bar{6}$	0,3000	0,2000	0,1900	0,1890	0,1677	0,1652	0,1666

Der Zusammenhang lässt sich auch graphisch veranschaulichen. Abbildung 10.1 zeigt vier Säulendiagramme. Links oben sind die relativen Häufigkeiten der Augenzahlen nach 10 Würfen dargestellt, rechts daneben die relativen Häufigkeiten nach 100 Würfen, links unten die relativen Häufigkeiten nach 1.000 Würfen und rechts unten die relativen Häufigkeiten nach 1.000.000 Würfen. Die Wahrscheinlichkeit P(A) ist gestrichelt eingezeichnet.

Da hier die relativen Häufigkeiten die Höhe der Säulen bestimmen und die Grundfläche der Säulen eine Einheit breit sind, entsprechen die Flächeninhalte genau der relativen Häufigkeit bzw. der Wahrscheinlichkeit. In der Summe ergeben die Flächen 1, was der Summe aller relativen Häufigkeiten entspricht.

Abbildung 10.1: Simulation des Werfens eines Würfels



Gestrichelte Linien: $P(A) = 1/6 = 0{,}1\overline{6}$

Im Prinzip haben wir mit diesem `Experiment I` das Ziehen einer Stichprobe mit Zurücklegen simuliert. Die Stichprobengröße variierte von 10 über 50, 100, 1.000, 10.000, 100.000 bis zu 1.000.000. Für jede Stichprobe haben wir die relativen Häufigkeiten – die Anteile – der Ausprägungen eines diskreten Merkmals berechnet. Der „wahre" Anteil in der Grundgesamtheit (θ, sprich: theta) entspricht der Wahrscheinlichkeit eines Ereignisses $P(A)$. Wie wir gesehen haben, konvergiert die relative Häufigkeit einer Merkmalsausprägung in einer Stichprobe mit wachsendem Stichprobenumfang gegen den Anteil der Grundgesamtheit.

Allerdings kann man die Vergrößerung der Stichprobe nicht wie die Anzahl der Würfe beim Würfel-Experiment unendlich oft ausdehnen, denn irgendwann stimmt die Größe der Stichprobe mit der der Grundgesamtheit überein. Diese Verletzung der Voraussetzung des Bernoulli-Theorems – eine unendliche Wiederholung des Zufallsexperiments entspricht einer Vergrößerung der Stichprobe ins Unendliche – ist aber nur dann relevant, wenn die Grundgesamtheit nicht besonders groß ist oder umgekehrt die zu ziehende Stichprobe sehr groß ist. Wenn die Grundgesamtheit die Stichprobe um den Faktor 100 übersteigt, fällt die Verletzung dieser Voraussetzung schon nicht mehr ins Gewicht.

Weil die relativen Häufigkeiten sich den Wahrscheinlichkeiten nähern, aber nicht identisch sind, wäre es wünschenswert zu wissen, wie die relativen Häufigkeiten von den Wahrscheinlichkeiten abweichen. Dazu führen wir ein neues Experiment durch. Wir vereinfachen `Experiment II` im Vergleich zu `Experiment I` dadurch, dass nur die relative Häufigkeit des Auftretens einer einzigen Augenzahl, nämlich der 6, notiert wird. Die Wahrscheinlichkeit dafür beträgt $P(A) = 1/6 = 0{,}1\bar{6}$. Die Gegenwahrscheinlichkeit $P(\bar{A})$ – also die Wahrscheinlichkeit *keine* 6 zu werfen – entspricht $P(\bar{A}) = 1 - P(A) = 1 - 0{,}1\bar{6} = 0{,}8\bar{3}$.

Ein Zufallsexperiment bei dem nur zwei Ausgänge (Ereignis und Gegenereignis) möglich sind, wird als *Bernoulli-Experiment* bezeichnet. Eine Reihe nacheinander durchgeführter Bernoulli-Experimente (z. B. mehrmaliges Würfeln) nennt man Bernoulli-Kette. `Experiment II` besteht darin, *den Würfel 100-mal zu werfen und die Häufigkeit des Auftretens der 6 festzuhalten.* Dieser Versuch wird *zehn* Mal wiederholt. Die Wahrscheinlichkeit für das Auftreten der Zahl 6 ist $P(A) = 0{,}1\bar{6}$. Theoretisch müsste die 6 nach 100 Würfen ca. 16 oder 17-mal fallen, was einem prozentualen Anteil von $16{,}\bar{6}\,\%$ entspricht.

In Tabelle 10.2 sind die *Anteile* der Augenzahl 6 bei 100 Würfen für die zehn Versuche notiert. Wie man sieht, liegt der beobachtete Wert nur im 7. Versuch in der Nähe des theoretisch erwarteten Wertes von $16{,}\bar{6}\,\%$.

Tabelle 10.2: Anteilswerte der Zahl 6 bei 100 Würfen

Versuch Nr.	1	2	3	4	5	6	7	8	9	10
Anteil in %	20	19	20	20	20	15	17	13	13	18

Alle anderen Werte weichen mehr oder weniger vom erwarteten Wert ab. Was können wir daraus folgern? Wir wissen aufgrund des Bernoulli-Theorems, dass bei einer sehr großen Zahl von Würfen (z. B. eine 1.000.000-mal) die relative Häufigkeit der Wahrscheinlichkeit nahezu entspricht. Würfeln wir nicht so häufig, dann weichen die relativen Häufigkeiten stärker von der Wahrscheinlichkeit ab. Die Lösung des Problems besteht darin, dass Stichprobenkennwerte nicht „irgendwie“ vom erwarteten Wert abweichen. Die Abweichung kann durch Wahrscheinlichkeitsverteilungen angegeben werden.

10.2 Häufigkeiten und Anteile in Stichproben

10.2.1 Binomialverteilung

Betrachten wir zunächst eine Häufigkeitsauszählung des `Experiments II`, wenn wir den Versuch nicht 10-mal, sondern 1.000-mal durchführen. Wir notieren wieder die Häufigkeit, mit der bei jeweils 100 Würfen die Zahl 6 fällt. Dies entspricht dem Ziehen von 1.000 Stichproben des Umfangs $n = 100$ mit Zurücklegen. Theoretisch kann die 6 bei jedem dieser 1.000 Versuchsdurchführungen zwischen 0 und 100-mal fallen.

Wie wir Tabelle 10.3 entnehmen können, kommen jedoch nur bestimmte Häufigkeiten vor, und manche Werte kommen wesentlich öfter vor als andere. So kann man der vierten Zeile der Tabelle entnehmen, dass in 18 von den insgesamt 1.000 Durchführungen (=1,8 % der Experimente) die Zahl 6 bei 100 Würfen genau 9-mal fiel, dies entspricht einem Anteil der Zahl 6 von 9 %.

Man sieht, dass Anteile, die relativ weit vom erwarteten Wert ($16{,}\bar{6}$ % = $0{,}1\bar{6} \cdot 100$) entfernt sind, nur selten oder nie vorkommen, während Häufigkeiten und Anteilswerte in der Nähe des erwarteten Wertes liegen, sehr häufig auftreten. Am häufigsten, nämlich in 106 der 1.000 Versuchsdurchführungen (10,6 %), fiel die Augenzahl 6 bei 100 Würfen 17-mal, d. h. bei 17 % der 100 Würfe. Wie man an der kumulierten Häufigkeitsverteilung in der letzten Spalte von Tabelle 10.3 ablesen kann, liegen 51 % der Anteilswerte unterhalb von 17 % und entsprechend 49 % der Anteilswerte über diesem Wert. Durch ein Histogramm kann man die Verteilung graphisch veranschaulichen (Abbildung 10.2 auf Seite 230).

Die Wahrscheinlichkeitsverteilung der Zufallsvariablen „Häufigkeit bzw. Anteil der 6 bei 100-maligem Würfeln“ ist die **Binomialverteilung**. Die Wahrscheinlichkeitsverteilung ist in Abbildung 10.2 als gestrichelte Linie eingezeichnet. Die Gleichung lautet:

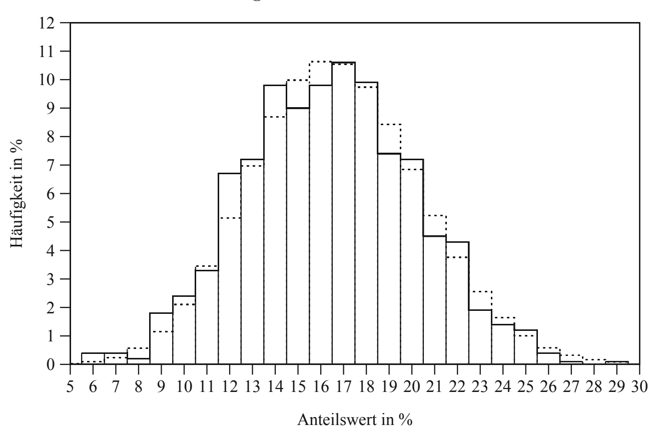
$$f_B(x|n;\theta) = \binom{n}{x} \cdot \theta^x \cdot (1-\theta)^{n-x}, \qquad \text{für} \quad x = 0, 1, 2, \ldots, n\,. \quad (10.2)$$

Die linke Seite dieser Gleichung ist die Bezeichnung für eine Binomialverteilung. Da es eine Wahrscheinlichkeitsfunktion ist, wird der Ausdruck $f_B(x|n;\theta)$ als „Wahrscheinlichkeit für x unter der Bedingung, dass

Tabelle 10.3: Anteilswerte der Zahl 6 bei 100 Würfen und 1.000 Wiederholungen

	Häufigkeit		kum. Häufigkeit	
Anteil	absolut	in %	absolut	in %
6 %	4	0,4	4	0,4
7 %	4	0,4	8	0,8
8 %	2	0,2	10	1,0
9 %	18	1,8	28	2,8
10 %	24	2,4	52	5,2
11 %	33	3,3	85	8,5
12 %	67	6,7	152	15,2
13 %	72	7,2	224	22,4
14 %	98	9,8	322	32,2
15 %	90	9,0	412	41,2
16 %	98	9,8	510	51,0
17 %	106	10,6	616	61,6
18 %	99	9,9	715	71,5
19 %	74	7,4	789	78,9
20 %	72	7,2	861	86,1
21 %	45	4,5	906	90,6
22 %	43	4,3	949	94,9
23 %	19	1,9	968	96,8
24 %	14	1,4	982	98,2
25 %	12	1,2	994	99,4
26 %	4	0,4	998	99,8
27 %	1	0,1	999	99,9
29 %	1	0,1	1000	100,0

Abbildung 10.2: Anteilswerte der Zahl 6 bei 100 Würfen und 1.000 Wiederholungen



n und θ zutrifft“ gelesen. n ist die Zahl der Wiederholungen des Bernoulli-Experimentes (hier: 100 Würfe). θ (theta) $= P(A)$ ist die Wahrscheinlichkeit des interessierenden Ereignisses (hier: 1/6 für das Auftreten der 6). n und θ sind die Parameter der Binomialverteilung. x gibt die Häufigkeit an, mit der das interessierende Ereignis eintritt.

Die Gleichung kann leicht hergeleitet werden. Dazu fragen wir, wie wahrscheinlich es ist, dass die Zahl 6 bei 100 Würfen 20-mal auftritt. A kennzeichnet das interessierende Ereignis – das Auftreten einer 6, $\bar{A}$ das Gegenereignis – das Auftreten einer anderen Zahl. Wir betrachten zunächst eine konkrete Stichprobe

$$(\underbrace{A, A, \ldots, A}_{x\text{-mal}}, \underbrace{\bar{A}, \bar{A}, \ldots, \bar{A}}_{(n-x)\text{-mal}}) .$$

Die ersten x Elemente der Stichprobe sind das interessierende Ereignis, die letzten $(n - x)$ Elemente das Gegenereignis. Beispielsweise könnte in den 20 ersten Würfen eine 6 auftreten, in den letzten 80 Würfen eine andere Zahl.

Die einzelnen Würfe sind voneinander unabhängig. Die Wahrscheinlichkeit dieser Reihenfolge lässt sich daher durch Multiplikation der Wahrscheinlichkeiten ermitteln, mit der die einzelnen Ereignisse auftreten (vgl. Gleichung 9.4, S. 204).

$$\underbrace{\theta \cdot \ldots \cdot \theta}_{x\text{-mal}} \cdot \underbrace{(1-\theta) \cdot \ldots \cdot (1-\theta)}_{(n-x)\text{-mal}} = \theta^x \cdot (1-\theta)^{(n-x)} \tag{10.3}$$

So ist die Wahrscheinlichkeit in den ersten 20 Würfen eine 6 und in den letzten 80 Würfen eine andere Augenzahl zu erzielen: $0{,}1\bar{6}^{20} \cdot (1 - 0{,}1\bar{6})^{100-20}$. Die Reihenfolge, in der die Zahl 6 auftritt, spielt keine Rolle für die Berechnung der Wahrscheinlichkeit.

Wenn wir danach fragen, mit welcher Wahrscheinlichkeit wir bei 100 Würfen 20-mal die Zahl 6 erzielen, dann müssen wir alle Stichproben (Reihenfolgen) berücksichtigen, in denen die Zahl 6 20-mal auftritt. So kann die Zahl 6 z. B. bei den ersten 10 Würfen und bei den letzten 10 Würfen auftreten. Insgesamt gibt es $\binom{n}{x}$ Möglichkeiten, x aus n Elementen auszuwählen (vgl. S. 206). In $\binom{100}{20}$ Stichproben tritt die Zahl 6 daher 20-mal auf. Jede dieser Stichproben hat die Realisierungswahrscheinlichkeit $0{,}1\bar{6}^{20} \cdot (1 - 0{,}1\bar{6})^{100-20}$.

Die Wahrscheinlichkeit, dass die Zahl 6 bei 100 Würfen 20-mal auftritt, ist daher:

$$\begin{aligned} f_B(20|100; 0{,}1\bar{6}) &= \binom{100}{20} \cdot 0{,}1\bar{6}^{20} \cdot (1 - 0{,}1\bar{6})^{100-20} \\ &= \frac{100!}{20! \cdot (100-20)!} \cdot 2{,}735 \cdot 10^{-16} \cdot 4{,}629 \cdot 10^{-7} \\ &= 5{,}359833704038 \cdot 10^{20} \cdot 1{,}266 \cdot 10^{-22} \\ &= 0{,}0679\,. \end{aligned}$$

Aus Tabelle 10.3 auf Seite 229 und Abbildung 10.2 geht hervor, dass tatsächlich in 7,2 % der 1.000 Experimente 20-mal die 6 geworfen wurde. Dieser Wert würde sich mit zunehmender Zahl von Experimenten immer

mehr dem theoretisch zu erwartenden Wert von 6,79 % annähern. Die erwarteten Anteilswerte der Zahl 6 bei 100 Würfen sind in Abbildung 10.2 mit einer gestrichelten Linie eingezeichnet.

Was bedeutet das für den Schluss von einer Stichprobe auf die Grundgesamtheit? Häufigkeiten und Anteile weichen in einfachen Zufallsstichproben vom Parameter der Grundgesamtheit ab. Sie tun dies jedoch nicht beliebig sondern in Form einer Binomialverteilung. Große Abweichungen sind sehr unwahrscheinlich, kleine Abweichungen wahrscheinlicher. Weil wir die Wahrscheinlichkeitsverteilung der Zufallsvariablen X und P – Häufigkeiten und Anteile in einfachen Zufallsstichproben mit Zurücklegen – kennen, können wir die Abweichungen der Stichprobenkennwerte vom Parameter der Grundgesamtheit berechnen.

Erwartungswert und Varianz

Auch Wahrscheinlichkeitsverteilungen lassen sich durch Maßzahlen beschreiben. In Kapitel 6 haben wir das arithmetische Mittel und die Varianz zur Beschreibung einer empirischen Verteilung angegeben. Analog dazu bezeichnen der **Erwartungswert** und die **Varianz** einer Zufallsvariablen die zentrale Lage und die Streuung einer Wahrscheinlichkeitsverteilung.

Erwartungswert und Varianz der Wahrscheinlichkeitsverteilung der *Häufigkeiten* können einfach durch folgende Formeln ermittelt werden:

$$E(X) = n \cdot \theta \quad \text{und} \tag{10.4}$$

$$Var(X) = n \cdot \theta \cdot (1 - \theta). \tag{10.5}$$

Der Standardfehler gibt die Standardabweichung der Wahrscheinlichkeitsverteilung an und berechnet sich als $\sqrt{Var(X)}$. In unserem Beispiel resultiert nach Gleichung 10.4

$$\begin{aligned} E(X) &= n \cdot \theta \\ &= 100 \cdot 0{,}1\bar{6} \\ &= 16{,}\bar{6} \end{aligned}$$

und nach Gleichung 10.5

$$\begin{aligned} Var(X) &= n \cdot \theta(1-\theta) \\ &= 100 \cdot 0{,}1\bar{6} \cdot 0{,}8\bar{3} \\ &= 13{,}\bar{8}\,. \end{aligned}$$

Wir erwarten also bei einem Experiment, dass von 100 Würfen durchschnittlich 16,$\bar{6}$-mal die 6 fällt. Die Varianz dieses Wertes bei allen Experimenten beträgt 13,$\bar{8}$, die Standardabweichung $\sqrt{13{,}\bar{8}} = 3,7$.

Wir haben die Häufigkeiten des Auftretens eines bestimmten Ereignisses oben bereits in Anteile umgerechnet. Erwartungswert und Varianz der Verteilung der Zufallsvariablen P – *Anteilswerte* – ergeben sich nach

$$E(P) = \frac{1}{n} \cdot E(X) = \theta \quad \text{und} \tag{10.6}$$

$$Var(P) = \frac{1}{n^2} \cdot Var(X) = \frac{\theta(1-\theta)}{n}\,. \tag{10.7}$$

Die Standardabweichung der Anteilswerteverteilung berechnet sich aus der Quadratwurzel der Varianz

$$\sigma_p = \sqrt{\frac{\theta(1-\theta)}{n}}\,. \tag{10.8}$$

Die Standardabweichung von Zufallsvariablen, die zur Schätzung von Parametern der Grundgesamtheit verwandt werden, nennt man *Standardfehler* bzw. Standardschätzfehler. σ_p ist der Standardfehler des Anteilswertes. Standardfehler messen die Größe des Stichprobenfehlers. Der Standardfehler des Anteils wird umso kleiner, je größer der Stichprobenumfang n ist, wie man an Formel 10.8 sehen kann.

$\theta(1-\theta)$ gibt die Varianz des Anteils in der Grundgesamtheit an, wenn das interessierende Ereignis A mit 1 und das Gegenereignis $\bar{A}$ mit 0 kodiert ist.

Die Varianz des Anteils in der Grundgesamtheit ist am größten, wenn $\theta = 0,5$ ist. Je weiter θ von 0,5 entfernt ist, je kleiner die Varianz des Anteils in der Grundgesamtheit. Der Standardfehler des Anteils in Stichproben steigt mit der Varianz des Anteils in der Grundgesamtheit.

Im Beispiel erhält man für $n = 100$ Würfe einen Erwartungswert des Anteils von

$$\begin{aligned} E(P) &= \frac{1}{n}E(X) \\ &= \frac{1}{100} \cdot 16{,}\bar{6} \\ &= 0{,}1\bar{6}\,, \end{aligned}$$

eine Varianz von

$$\begin{aligned} Var(P) &= \frac{\theta(1-\theta)}{100} \\ &= \frac{0{,}1\bar{6}(1-0{,}1\bar{6})}{100} \quad = 0,13\bar{8} \end{aligned}$$

und einen Standardfehler von

$\sigma_p = \sqrt{0,13\bar{8}} = 0{,}37.$

10.2.2 Hypergeometrische Verteilung

Würfelt man mehrmals hintereinander, dann entspricht das dem Ziehen einer Stichprobe mit Zurücklegen. Für Umfragen zieht man Stichproben *ohne* Zurücklegen. Die Wahrscheinlichkeitsverteilung für Häufigkeiten und Anteile in einfachen Zufallsstichproben ohne Zurücklegen ist die hypergeometrische Verteilung. Die Erwartungswerte sind bei der hypergeometrischen Verteilung und der Binomialverteilung identisch. Bei der Berechnung der Varianz wird allerdings noch mit dem Faktor $(N-n)/(N-1)$ multipliziert. Für Häufigkeiten resultiert

$$Var(X) = n \cdot \theta \cdot (1 - \theta) \cdot \frac{N - n}{N - 1} \tag{10.9}$$

und für Anteile

$$Var(P) = \frac{\theta \cdot (1 - \theta)}{n} \cdot \frac{N - n}{N - 1}. \tag{10.10}$$

Für eine gegebene Stichprobengröße n nähert sich $(N - n)/(N - 1)$ mit zunehmender Größe der Grundgesamtheit N dem Wert 1 an. Ist der Quotient aus dem Umfang der Grundgesamtheit und dem Stichprobenumfang größer als 20, $N/n > 20$, kann der Korrekturfaktor $(N - n)/(N - 1)$ vernachlässigt werden. Dies ist in der Praxis häufig der Fall.

10.3 Stichprobenmittelwerte

Unser Beispiel bezog sich bisher auf den Fall, dass wir es mit einer diskreten Zufallsvariablen zu tun haben. Eine stetige Zufallsvariable wäre das Merkmal Alter in einfachen Zufallsstichproben. Stetige Merkmale haben nicht abzählbar viele Ausprägungen. Dagegen wäre das Merkmal Geschlecht eine diskrete Zufallsvariable, da es nur die Ausprägungen „Mann“ und „Frau“ besitzt.

Im Kapitel 10.2.1 haben wir ein Experiment (100-mal Würfeln) mehrmals wiederholt. In der Praxis entspricht dies dem Ziehen mehrerer Stichproben der Stichprobengröße 100. Bei jedem Experiment haben wir festgehalten, wie oft die 6 gefallen ist, was der Feststellung der Häufigkeit - und daraus abgeleitet - des Anteilswertes in einer Stichprobe entspricht.

Wir führen nun ein neues Experiment durch. `Experiment III` besteht darin, aus der Altersverteilung der bundesdeutschen Bevölkerung im Jahr 1974 einfache Zufallsstichproben (ohne Zurücklegen) mit jeweils 1.000 Befragten zu ziehen. Insgesamt ziehen wir 1.000 verschiedene Stichproben. Für jede Stichprobe berechnen wir den Altersdurchschnitt $\bar{x}$. Die Stichprobenziehung simulieren wir mit dem Programm ALTMIHI aus GSTAT (vgl. Böker 1993). GSTAT enthält die Altersverteilung der bundesdeutschen Bevölkerung im Jahr 1974. Der Altersdurchschnitt in der bundesdeutschen

Bevölkerung lag 1974 bei $\mu = 37{,}27$ Jahre, die Varianz betrug $\sigma^2 = 504,45$. μ (sprich: mü) und σ^2 (sprich: sigma-Quadrat) kennzeichnen den Mittelwert und die Varianz der Grundgesamtheit.

Häufigkeitsauszählungen stetiger Variablen werden dargestellt, indem man die Merkmalsausprägungen in Intervalle zusammenfasst und die Häufigkeit der Werte in diesen Intervallen berichtet. Die Verteilung der Altersdurchschnitte wird in Intervalle der Breite 0,1 eingeteilt. Jedes Intervall hat eine untere und obere Grenze, z. B. reicht das erste Intervall von 34,75 bis 34,85 Jahre. Statt der Intervallgrenzen kann als Kategorie auch einfach die Intervallmitte angegeben werden, wie das in der folgenden Tabelle 10.4 zu sehen ist. Das erste Intervall hat z. B. die Mitte 34,8 Jahre. In dieses Intervall fällt der Mittelwert einer einzigen Stichprobe, was bei 1.000 Stichproben zur relativen Häufigkeit 0,001 führt (bzw. zur prozentualen Häufigkeit 0,1 %).

Man sieht, dass manche Altersdurchschnitte deutlich häufiger ermittelt werden als andere. Besonders häufig treten Stichproben auf, deren Altersdurchschnitt $\bar{x}$ nah am Wert der Grundgesamtheit μ liegt. Größere Abweichungen vom Parameter der Grundgesamtheit sind also auch hier, wie schon in Tabelle 10.3 auf Seite 229, selten, dagegen sind kleinere Abweichungen häufiger.

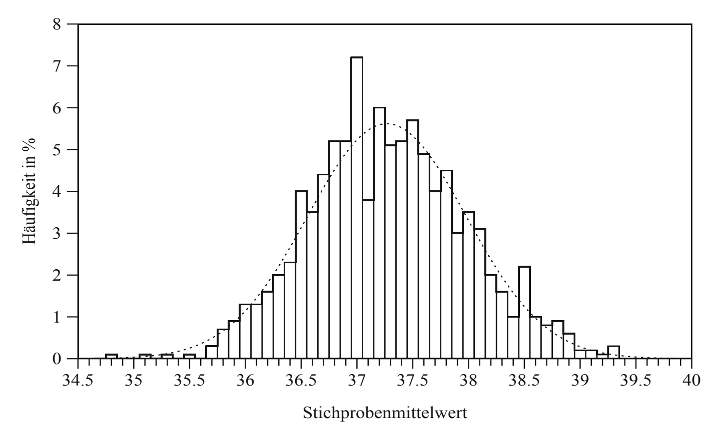
In Abbildung 10.3 ist die Verteilung der Altersdurchschnitte in den 1.000 Stichproben als Histogramm dargestellt. Die gestrichelte Linie in Abbildung 10.3 ist die **Wahrscheinlichkeitsfunktion der Normalverteilung**. Wie man sieht, ist die Verteilung der Altersdurchschnitte in den 1.000 Stichproben (Histogramm) annähernd normalverteilt (gestrichelte Linie).

Wenn wir viele Stichproben eines hinreichend großen Umfangs n ziehen, dann verteilen sich die arithmetischen Mittel dieser Stichprobenwerte normal. In der Praxis ist die Ziehung vieler Stichproben nicht notwendig. Mit dem *Zentralen Grenzwertsatz* lässt sich theoretisch begründen, dass die Zufallsvariable $\bar{X}$ – Mittelwerte in Stichproben – normalverteilt ist. Und zwar unabhängig von der Verteilung des Merkmals in der Grundgesamtheit. Wir werden auf die Verteilung von Stichprobenmittelwerten und den Zentralen Grenzwertsatz zurückkommen. Zunächst aber zur Normalverteilung.

Tabelle 10.4: Altersdurchschnitte bei 1.000 Stichproben der Größe 1.000

	Häufigkeit		kum. Häufigkeit	
Intervallmitte	absolut	in %	absolut	in %
34,8	1	0,10	1	0,10
35,1	1	0,10	2	0,20
35,3	1	0,10	3	0,30
35,5	1	0,10	4	0,40
35,7	3	0,30	7	0,70
35,8	7	0,70	14	1,40
35,9	9	0,90	23	2,30
36,0	13	1,30	36	3,60
36,1	13	1,30	49	4,90
36,2	16	1,60	65	6,50
36,3	20	2,00	85	8,50
36,4	23	2,30	108	10,80
36,5	40	4,00	148	14,80
36,6	35	3,50	183	18,30
36,7	44	4,40	227	22,70
36,8	52	5,20	279	27,90
36,9	52	5,20	331	33,10
37,0	72	7,20	403	40,30
37,1	38	3,80	441	44,10
37,2	60	6,00	501	50,10
37,3	51	5,10	552	55,20
37,4	52	5,20	604	60,40
37,5	57	5,70	661	66,10
37,6	49	4,90	710	71,00
37,7	40	4,00	750	75,00
37,8	45	4,50	795	79,50
37,9	30	3,00	825	82,50
38,0	35	3,50	860	86,00
38,1	31	3,10	891	89,10
38,2	20	2,00	911	91,10
38,3	16	1,60	927	92,70
38,4	10	1,00	937	93,70
38,5	22	2,20	959	95,90
38,6	10	1,00	969	96,90
38,7	8	0,80	977	97,70
38,8	9	0,90	986	98,60
38,9	6	0,60	992	99,20
39,0	2	0,20	994	99,40
39,1	2	0,20	996	99,60
39,2	1	0,10	997	99,70
39,3	3	0,30	1000	100,00

Abbildung 10.3: Altersdurchschnitte bei 1.000 Stichproben der Größe 1.000



10.3.1 Normalverteilung und Standardnormalverteilung

Die Normalverteilung trägt auch die Namen „Gauß'sche Normalverteilung" – nach ihrem „Mitbegründer" Carl Friedrich Gauß – oder „Glockenkurve" – wegen ihres charakteristischen, an eine Glocke erinnernden Verlaufs. Die Normalverteilung ist bedeutsam

1. als Verteilung empirischer Merkmale,
2. als Verteilung von Kennwerten in Stichproben und
3. als Approximation anderer theoretischer Verteilungen.

Bei der Körpergröße (Abbildung 6.1, S. 129) oder den Mathematikkenntnissen (Abbildung 8.2, S. 178) handelt es sich näherungsweise um normalverteilte Merkmale. Empirische Merkmale sind in der Regel jedoch nicht normalverteilt. Die Bedeutung der Normalverteilung in der Statistik resultiert vor allem aus den beiden letztgenannten Punkten. Sie gibt die Verteilung von Stichprobenmittelwerten an, wie wir im letzten Abschnitt gesehen haben. Außerdem können viele Verteilungen durch die Normalverteilung angenähert werden. Unter bestimmten Voraussetzungen geht beispielsweise die Binomialverteilung in eine Normalverteilung über, wie

wir am Ende des Kapitels sehen werden (vgl. die Übersicht bei Bleymüller et al. 2004, Kapitel 11).

Die Normalverteilungsfunktion für ein empirisches Merkmal in der Stichprobe lautet:

$$f_N(x|\bar{x}; s^2) = \frac{1}{s \cdot \sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\bar{x}}{s})^2} . \qquad (10.11)$$

Ihre beiden Parameter $\bar{x}$ – das ist das arithmetische Mittel der Verteilung – und s^2 – das ist die Varianz der Verteilung – bestimmen dabei die Lage und Breite der Kurve. Um eine Normalverteilung zu charakterisieren, reicht die Angabe des Mittelwertes und der Varianz daher aus. Aus diesem Grunde werden Normalverteilungen meistens mit $N(\bar{x}|s^2)$ bezeichnet. Bezieht sich die Normalverteilung auf ein Merkmal der Grundgesamtheit, dann werden die griechischen Bezeichnungen für das arithmetische Mittel und die Varianz – μ und σ^2 – verwendet: $N(\mu|\sigma^2)$.

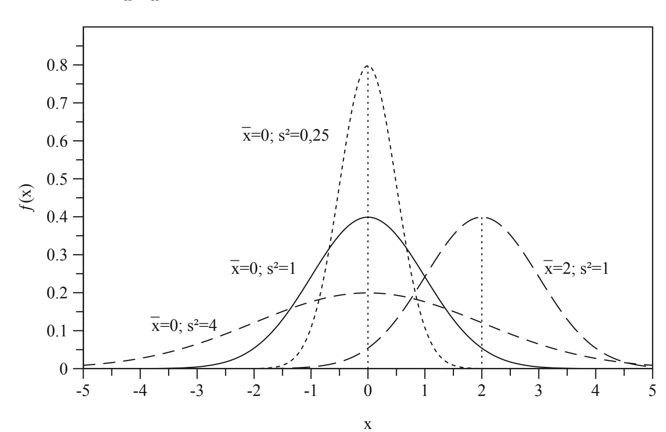
Die Funktion gibt die Wahrscheinlichkeitsdichte an. Zur Verdeutlichung sind in Abbildung 10.4 auf der nächsten Seite mehrere Normalverteilungen mit verschiedenen Parametern $\bar{x}$ und s^2 dargestellt.

Man kann erkennen, dass die Verteilung mit größer werdender Varianz s^2 breiter und bei kleiner werdender Varianz s^2 schmaler wird. Wird der Mittelwert $\bar{x}$ größer, so verschiebt sich die Verteilung auf der x-Achse nach rechts, wird der Mittelwert $\bar{x}$ kleiner, so verschiebt sie sich nach links.

Die Normalverteilung hat mehrere Eigenschaften, die man sich bei der Anwendung in der Inferenzstatistik zunutze machen kann:

- Sie ist symmetrisch und eingipflig, wobei ihr Maximum bei $\bar{x}$ liegt. Arithmetisches Mittel, Modalwert und Median sind aus diesem Grund identisch.
- Sie nähert sich asymptotisch der x-Achse, d. h. dem Wert null, wenn x gegen $+\infty$ oder $-\infty$ strebt. Sie wird jedoch nie gleich null (auch wenn es in der Abbildung so aussehen sollte).
- Ihre Wendepunkte - die steilsten Stellen – liegen bei $\bar{x} - s$ und $\bar{x} + s$.
- Da die Verteilung symmetrisch ist, befinden sich 50 % der Fläche links von $\bar{x}$ und 50 % rechts von $\bar{x}$.

Abbildung 10.4: Normalverteilungen mit verschiedenen Parametern $\bar{x}$ und s^2



Die Fläche unterhalb der Normalverteilung gibt an, wie viele x-Werte sich in einem bestimmten Bereich der Verteilung befinden. Zwischen $x_1 = -\infty$ und $x_2 = +\infty$ befinden sich alle x-Werte, also 100 %, die dazugehörige Fläche beträgt demnach 1. Wir können beliebige Flächen unter der Normalverteilung bestimmen. Am einfachsten geschieht dies, indem wir uns der Standardnormalverteilung bedienen.

Die Standardnormalverteilung ist eine Normalverteilung, deren Mittelwert null und deren Varianz eins ist. Gleichung 10.11 vereinfacht sich zur **Dichtefunktion der Standardnormalverteilung**:

$$\phi(z) = f_N(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \,. \tag{10.12}$$

Die Werte der Standardnormalverteilung werden als **z-Werte** bezeichnet, da man die Werte jeder beliebigen Normalverteilung mittels einer **z-Transformation** in eine Standardnormalverteilung überführen kann. Das Besondere der Standardnormalverteilung besteht darin, dass die Werte der

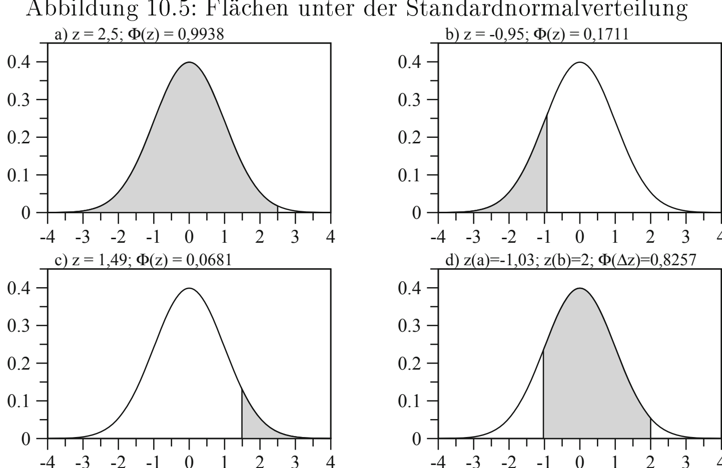
Verteilungsfunktion in vielen Statistikbüchern in tabellierter Form vorliegen (vgl. Anhang A, S. 308). Die Verteilungsfunktion $F_N(z)$ der Standardnormalverteilung gibt die Wahrscheinlichkeit dafür an, dass die Zufallsvariable Z höchstens den Wert z annimmt. Diese entspricht der Fläche, die *links* von einem z-Wert liegt. Zur Kennzeichnung der Verteilungsfunktion verwendet man den griechischen Großbuchstaben Φ (Phi). Wie man an der Tabelle sieht, braucht man nur z-Werte von -3 bis $+3$ auszuweisen, da die Fläche für z-Werte kleiner als -3 nahezu 0 ist, und für z-Werte größer als $+3$ nahezu 1 bzw. 100 %.

Beispielsweise geht aus der z-Tabelle im Anhang A hervor, dass links vom z-Wert 0 die Fläche 0,5 bzw. 50 % liegt. Da es sich um eine zum Mittelwert 0 symmetrische Verteilung handelt, ist die Fläche, die sich links vom Wert 0 befindet genauso groß wie die Fläche rechts vom Wert 0. Links vom z-Wert $+2{,}5$ befinden sich 99,38 % der Fläche, links vom z-Wert $-0{,}95$ sind es 17,11 %. Wenn man wissen möchte, welcher Flächenanteil sich *rechts* vom z-Wert befindet, kann man sich die Tatsache zunutze machen, dass sich unter der gesamten Verteilung die Fläche 1 bzw. 100 % befindet. Wenn links von einem z-Wert die Fläche Φ ist, dann befindet sich rechts vom selben Wert die Fläche $1 - \Phi$. *Rechts* vom z-Wert $+1{,}49$ liegt also die Fläche $1 - 0{,}9319 = 0{,}0681$, d. h. 6,81 % der Gesamtfläche.

Ganz ähnlich lässt sich auch verfahren, wenn man wissen möchte, wie groß die Fläche *innerhalb bestimmter Grenzen* bzw. *innerhalb eines Intervalles* ist. Da die Tabelle immer die Fläche ausweist, die links von einem z-Wert liegt, muss man, um ein Intervall zwischen zwei Werten zu erhalten, von der Fläche, die links vom größeren z-Wert (z_2) liegt, die Fläche, die links vom kleineren z-Wert (z_1) liegt, subtrahieren: $\Phi(\Delta z) = \Phi_{z_2} - \Phi_{z_1}$. Zwischen $-1{,}03$ und $+2$ befinden sich also $\Phi_{z_2} - \Phi_{z_1} = 0{,}9772 - 0{,}1515 = 0{,}8257 = 82{,}57\,\%$ der Fläche.

Die Flächenberechnung ist in den Abbildungen 10.5 a) bis 10.5 d) visualisiert. Abbildung 10.5 a) zeigt die Fläche links vom z-Wert $+2{,}5$, Abbildung 10.5 b) links von $-0{,}95$. Abbildung 10.5 c) zeigt die Fläche rechts von $+1{,}49$ und Abbildung 10.5 d) schließlich die Fläche zwischen den z-Werten $-1{,}03$ und $+2$.

Abbildung 10.5: Flächen unter der Standardnormalverteilung



Für die Intervalle um den Mittelwert (-1;1), (-2;2) und (-3;3) lassen sich folgende Flächen festhalten:

1. Zwischen -1 und $+1$ liegen 68,27 % der Fläche bzw. der z-Werte.
2. Zwischen -2 und $+2$ liegen 95,45 % der Fläche bzw. der z-Werte.
3. Zwischen -3 und $+3$ liegen 99,73 % der Fläche bzw. der z-Werte.

Da jede Normalverteilung durch eine **z-Transformation** in eine Standardnormalverteilung überführt werden kann, können wir auch für jede beliebige Normalverteilung mit Hilfe der z-Tabelle Flächenanteile bestimmen, wenn wir zuvor den entsprechenden x-Wert z-transformiert haben. Ein Wert einer beliebigen Verteilung wird z-transformiert (und damit standardisiert), indem von diesem Wert den Mittelwert der Verteilung subtrahiert und das Ergebnis durch die Standardabweichung dividiert wird.

$$z_i = \frac{x_i - \bar{x}}{s} \,. \tag{10.13}$$

Wenn jeder Wert einer beliebigen Normalverteilung z-transformiert wird, erhält man eine Standardnormalverteilung mit den Parametern $\bar{x} = 0$ und

$s^2 = s = 1$. Ihre x_i-Werte sind jetzt z_i-Werte. Sie wird deshalb auch „z-Verteilung“ genannt. Auf dem umgekehrten Weg kann jede z-Verteilung in eine beliebige Verteilung mit den Parametern $\bar{x}$ und s^2 überführt werden:

$$x_i = \bar{x} + z_i \cdot s \tag{10.14}$$

Da auch jede *Normalverteilung* über die Umkehrung der z-Transformation gemäß Gleichung (10.14) aus der z-Verteilung ableitbar ist, lässt sich anhand der tabellierten z-Werte die **Flächenberechnung für jede Normalverteilung** durchführen.

Um herauszufinden, *wie viel Prozent der Fläche bei einer beliebigen Normalverteilung zwischen zwei x-Werten liegt*, z-standardisiert man zunächst die beiden x-Werte, um dann die Flächen für die standardisierten Werte aus der z-Tabelle abzulesen:

$$\begin{aligned}\Phi(\Delta x) &= \Phi_{x_2} - \Phi_{x_1} \\ &= \Phi_{(x_2-\bar{x})/s} - \Phi_{(x_1-\bar{x})/s} \\ &= \Phi_{z_2} - \Phi_{z_1} \,.\end{aligned}$$

Und auf ein Beispiel angewendet: In einer Normalverteilung mit dem Mittelwert $\bar{x} = 3$ und einer Standardabweichung von $s = 4$ soll die Fläche zwischen den Werten $x_1 = 2$ und $x_2 = 5$ berechnet werden:

$$\begin{aligned}\Phi(\Delta x) &= \Phi_5 - \Phi_2 \\ &= \Phi_{(5-3)/4} - \Phi_{(2-3)/4} \\ &= \Phi_{0{,}5} - \Phi_{-0{,}25} \\ &= 0{,}6915 - 0{,}4013 \\ &= 0{,}2902 = 29{,}02\,\% \,.\end{aligned}$$

Zwischen den beiden x-Werten 2 und 5 liegen in einer Normalverteilung mit dem Mittelwert 3 und der Standardabweichung 2 also 29,02 % der Werte.

Mit Hilfe der z-Standardisierung kann auch abgeleitet werden:

1. Zwischen $\bar{x} - 1 \cdot s$ und $\bar{x} + 1 \cdot s$ liegen 68,27 % der Fläche.
2. Zwischen $\bar{x} - 2 \cdot s$ und $\bar{x} + 2 \cdot s$ liegen 95,45 % der Fläche.
3. Zwischen $\bar{x} - 3 \cdot s$ und $\bar{x} + 3 \cdot s$ liegen 99,73 % der Fläche.

10.3.2 Die Verteilung der Stichprobenmittelwerte

Aus Abbildung 10.3 auf Seite 238 ging hervor, dass sich Mittelwerte aus mehreren Stichproben normalverteilen. Da es sich um Stichproben*mittelwerte* handelt, wurden in Abbildung 10.3 auf der x-Achse nicht x-Werte, sondern $\bar{x}$-Werte abgetragen. Die Stichprobenmittelwerte $\bar{X}$ sind Realisationen des Zufallsexperimentes Ziehen einer Zufallsstichprobe mit Zurücklegen.

Erwartungswert und Varianz

Das arithmetische Mittel – der Erwartungswert – der Stichprobenmittelwerteverteilung entspricht dem arithmetischen Mittelwert der Grundgesamheit μ:

$$E(\bar{X}) = \mu \,. \tag{10.15}$$

Die Varianz der Verteilung der Stichprobenmittelwerte lässt sich durch

$$Var(\bar{X}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{10.16}$$

berechnen.

Die *Standardabweichung* der Verteilung der Stichprobenmittelwerte $\sigma_{\bar{x}}$ berechnet sich als

$$\sigma_{\bar{x}} = \sqrt{\sigma_{\bar{x}}^2} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \,. \tag{10.17}$$

Sie wird als **Standardfehler des Mittelwerts** (auch: Standardschätzfehler) bezeichnet, um sie von der Standardabweichung des Merkmals in

der Grundgesamtheit (oder einer einzigen Stichprobe) zu unterscheiden. Wenn von Standardfehler gesprochen wird, dann ist immer die Breite einer Stichprobenkennwerteverteilung gemeint.

Der Standardfehler des Mittelwerts ist von zwei Faktoren abhängig: Zum einen von der Varianz des Merkmals in der Grundgesamtheit, σ^2. Je stärker ein Merkmal in der Grundgesamtheit streut, desto mehr werden auch die Mittelwerte des Merkmals in verschiedenen Stichproben voneinander abweichen. Zum anderen spielt die Stichprobengröße eine entscheidende Rolle: Je größer der Umfang der gezogenen Stichproben, umso weniger weichen diese vom Parameter der Grundgesamtheit μ ab. Die Verteilung der Stichprobenmittelwerte ist bei größerem n schmaler.

Gleichungen 10.15, 10.16 und 10.17 gelten für Stichproben, die mit Zurücklegen gezogen wurden und für Stichproben *ohne* Zurücklegen, in denen der Umfang der Grundgesamtheit N mindestens das 20-fache des Umfangs der Stichprobe entspricht, $\frac{N}{n} > 20$. Dies ist im Beispiel – Stichproben des Umfangs $n = 1.000$ aus der bundesdeutschen Bevölkerung – der Fall.

Der Erwartungswert des Durchschnittsalters in den Stichproben vom Umfang $n = 1.000$ ist

$$E(\bar{X}) = \mu = 37,27\,\text{Jahre}\,.$$

Varianz und Standardfehler des Mittelwertes betragen

$$\begin{aligned}\sigma^2_{\bar{x}} &= \frac{\sigma^2}{n}\\ &= \frac{504,45}{1000}\\ &= 0,506 \quad \text{und}\end{aligned}$$

$$\begin{aligned}\sigma_{\bar{x}} &= \sqrt{\frac{\sigma^2}{n}}\\ &= \sqrt{\frac{504,45}{1000}}\\ &= 0,71\,.\end{aligned}$$

Wir erwarten also durchschnittlich einen Stichprobenmittelwert, der dem Parameter der Grundgesamtheit entspricht, nämlich 37,3 Jahre. Der Standardfehler des Mittelwertes beträgt 0,71 Jahre.

Ist in Stichproben ohne Zurücklegen $\frac{N}{n} \leq 20$, dann muss die Varianz mit dem Korrekturfaktor für endliche Grundgesamtheiten multipliziert werden, den wir bereits in Kapitel 10.2.2 kennen gelernt haben.

Die Formeln für Varianz und Standardfehler beim Ziehen *ohne* Zurücklegen und $N/n \leq 20$ sind

$$Var(\bar{X}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} \quad \text{und} \tag{10.18}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}\,. \tag{10.19}$$

Wir werden den Korrekturfaktor für endliche Gesamtheiten in den nächsten Kapiteln vernachlässigen, weil N/n bei bevölkerungsweiten Umfragen größer als 20 ist. Immer wenn $N/n \leq 20$ bzw. $n/N \geq 0{,}05$ müssen die Standardfehler entsprechend korrigiert werden.

Wahrscheinlichkeitsfunktion

μ ist der Erwartungswert der Stichprobenmittelwerteverteilung, $\sigma_{\bar{x}}$ die Standardabweichung, die als Standardfehler bezeichnet wird. Die Gleichung der **Stichprobenmittelwerteverteilung** lautet:

$$f_N(\bar{x}|\mu;\sigma_{\bar{x}}^2) = \frac{1}{\sigma_{\bar{x}} \cdot \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}\right)^2}\,. \tag{10.20}$$

Anhand dieser Formel kann nun die Wahrscheinlichkeits*dichte* an der Stelle $\bar{x}$ bestimmt werden. Der Mittelwert der Altersverteilung in der bundesdeutschen Bevölkerung betrug 1974 exakt $\mu = 37{,}268$ Jahre, der Standardfehler des Mittelwerts $\sigma_{\bar{x}} = 0{,}71025$. Diese Werte werden als Parameter in die Gleichung (10.20) eingetragen. Die Wahrscheinlichkeitsdichte beträgt dann z. B. für $\bar{x}_i = 37{,}2$ Jahre

$$f_N(37{,}2|37{,}268; 0{,}71025^2) = \frac{1}{0{,}71025 \cdot \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{37{,}2-37{,}268}{0{,}71025}\right)^2}$$
$$= 0{,}5617 \cdot e^{-0{,}00458}$$
$$= 0{,}5591\,.$$

Dieser Wert ist aber nicht gleichbedeutend mit der Wahrscheinlichkeit des Punktes 37,2. Bei stetigen Verteilungen besitzt jeder einzelne der unendlich vielen Punkte die Wahrscheinlichkeit null, denn die Fläche – die ja die Wahrscheinlichkeit angibt – über Punkten ist null. In Abbildung 10.3 wurden zur Darstellung der Werte Intervalle der Breite 0,1 gewählt, somit muss der Wert der Wahrscheinlichkeits*dichte* mit der Intervallbreite 0,1 multipliziert werden, um ihn mit dem empirischen Wert vergleichen zu können: $0{,}5591 \cdot 0{,}1 = 0{,}05591$ bzw. 5,591 %. In Tabelle 10.4 (S. 237) sieht man, dass in 6 % der 1.000 simulierten Stichproben ein Altersdurchschnitt von 37,2 Jahren (Intervallmitte) ermittelt wurde. Der in den Simulationen ermittelte Wert von 6 % weicht von dem auf Basis der Normalverteilung theoretisch ermittelten Wert von 5,591 % geringfügig ab.

Flächen unter der Stichprobenmittelwerteverteilung

Die Flächenberechnung der Stichprobenmittelwerteverteilung kann wieder über die Tabelle der z-Verteilung erfolgen. Dazu werden die $\bar{x}$-Werte z-transformiert und die entsprechenden Flächenanteile aus der Tabelle übernommen. Das Vorgehen entspricht also dem Vorgehen bei einer beliebigen Normalverteilung. Allerdings lautet die Gleichung der z-Transformation für die Stichprobenmittelwerteverteilung

$$z_i = \frac{\bar{x}_i - \mu}{\sigma_{\bar{x}}} \tag{10.21}$$

und die Umkehrung

$$\bar{x}_i = \mu + z_i \cdot \sigma_{\bar{x}}\,. \tag{10.22}$$

$\bar{x}_i$ ist bei einer Stichprobenmittelwerteverteilung ein beliebiger Wert der Verteilung.

Analog zur Flächenberechnung unter der Normalverteilung lässt sich für die Stichprobenmittelwerteverteilung festhalten:

1. Zwischen $\mu - 1 \cdot \sigma_{\bar{x}}$ und $\mu + 1 \cdot \sigma_{\bar{x}}$ liegen 68,27 % der Stichprobenmittelwerte.
2. Zwischen $\mu - 2 \cdot \sigma_{\bar{x}}$ und $\mu + 2 \cdot \sigma_{\bar{x}}$ liegen 95,45 % der Stichprobenmittelwerte.
3. Zwischen $\mu - 3 \cdot \sigma_{\bar{x}}$ und $\mu + 3 \cdot \sigma_{\bar{x}}$ liegen 99,73 % der Stichprobenmittelwerte.

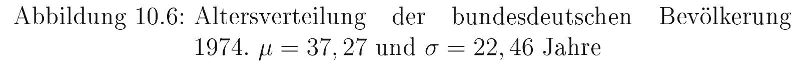
Im Beispiel liegen 95,45 % der Mittelwerte in Stichproben vom Umfang $n = 1000$ zwischen $37,3 - 2 \cdot 0,71 = 35,88$ und $37,3 + 2 \cdot 0,71 = 38,62$ Jahren.

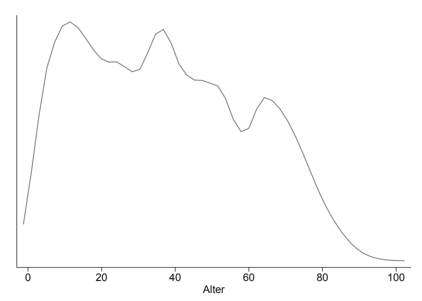
10.4 Der Zentrale Grenzwertsatz

Wir haben mit Hilfe von Simulationen demonstriert, dass sich Mittelwerte in Zufallsstichproben bei hinreichend großem Umfang n normalverteilen. Die theoretische Begründung liefert der Zentrale Grenzwertsatz. Der Zentrale Grenzwertsatz beinhaltet, dass eine Summe (und damit auch das arithmetische Mittel) identisch verteilter Zufallsvariablen mit zunehmendem Stichprobenumfang n gegen eine Normalverteilung konvergiert (vgl. Kühnel und Krebs 2007, 196 ff.).

Die Bedeutung des Zentralen Grenzwertsatzes liegt darin, dass sich die Stichprobenmittelwerte bei hinreichend großem Stichprobenumfang n auch dann normalverteilen, wenn das Merkmal in der Grundgesamtheit nicht normalverteilt ist. In Abbildung 10.6 ist die Altersverteilung der bundesdeutschen Bevölkerung 1974 dargestellt. Die Altersverteilung der Grundgesamtheit ist offensichtlich nicht normalverteilt. Dennoch nähern sich die Altersdurchschnitte in Stichproben vom Umfang $n = 1000$ einer Normalverteilung an (Abbildung 10.3, S. 238).

Bleibt die Frage, ab wann der Stichprobenumfang n groß genug ist, damit sich die Stichprobenmittelwerte normalverteilen. Die Antwort hängt von der Verteilung des Merkmals in der Grundgesamtheit ab. Bei Merkmalen, die in der Grundgesamtheit normalverteilt sind, verteilen sich die


Abbildung 10.6: Altersverteilung der bundesdeutschen Bevölkerung 1974. $\mu = 37,27$ und $\sigma = 22,46$ Jahre



arithmetischen Mittel in Zufallsstichproben unabhängig von der Größe der Stichproben immer normal. Bei sehr schiefen Verteilungen muss der Stichprobenumfang größer sein. Als Faustregel wird ein Stichprobenumfang von $n = 30$ angegeben, der notwendig ist, damit sich die Stichprobenmittelwerteverteilung einer Normalverteilung nähert.

Normalverteilung als Grenzverteilung für Häufigkeiten und Anteile

Wir haben gesehen, dass sich Häufigkeiten und Anteile binomial (bzw. hypergeometrisch) um den wahren Anteilswert der Grundgesamtheit verteilen. Bei großem Stichprobenumfang n ist die *Normalverteilung* auch die *Grenzverteilung für Häufigkeiten und Anteile*. Als hinreichend groß wird eine Stichprobe angesehen, wenn

$$n \cdot \theta \cdot (1 - \theta) > 9 \tag{10.23}$$

(Bleymüller et al. 2004, Kapitel 11.1). Alternativ wird gefordert, dass

$$n \cdot \frac{\theta}{1 - \theta} > 9 \quad \text{und} \quad n \cdot \frac{1 - \theta}{\theta} > 9 \tag{10.24}$$

(vgl. Kühnel und Krebs 2007, 204 ff.). Wird der Anteil der Stichprobe p zur Schätzung des Standardfehlers des Anteils verwendet (und nicht θ),

dann sollte der Stichprobenumfang zudem größer als 60 sein, $n > 60$ (vgl. Kühnel und Krebs 2007, 204 ff.). Darauf werden wir im nächsten Kapitel zurückkommen. Die Normalverteilung der Anteile hat die Parameter θ (Erwartungswert) und $\frac{\theta(1-\theta)}{n}$ (Varianz): $N(\theta; \frac{\theta \cdot (1-\theta)}{n})$. Bei Auswahlen ohne Zurücklegen und $\frac{N}{n} \leq 20$ ist die Varianz wieder mit dem Korrekturfaktor für endliche Grundgesamtheiten zu multiplizieren.

Grundgesamtheit - Stichprobe - Stichprobenkennwerte

Die Unterscheidung zwischen

1. der Verteilung eines Merkmals in der Grundgesamtheit,
2. der Verteilung eines Merkmals in der Stichprobe und
3. der Kennwerteverteilung

ist für das Verständnis der schließenden Statistik zentral. Zusammenfassend werden die drei Verteilungen am Beispiel der Altersverteilung deshalb noch einmal verdeutlicht. Die Verteilung eines Merkmals in der *Grundgesamtheit* ist in der Regel (und im Gegensatz zu den Beispielen aus diesem Kapitel) unbekannt. Wir bezeichnen das arithmetische Mittel eines Merkmals in der Grundgesamtheit mit μ und dessen Standardabweichung mit σ. Die Größe der Grundgesamtheit wird mit N bezeichnet.

Aus der Grundgesamtheit wird die *Stichprobe* vom Umfang n gezogen. Die Verteilung des Merkmals in der gezogenen Stichprobe ist bekannt. Das arithmetische Mittel und die Streuung eines Merkmals in der Stichprobe können wir aus den beobachteten Daten berechnen (Kapitel 6). Das arithmetische Mittel des Merkmals in der Stichprobe bezeichnen wir mit $\bar{x}$, die Standardabweichung mit s. Je größer der Stichprobenumfang n, desto ähnlicher wird die Verteilung des Merkmals in der Stichprobe der Verteilung des Merkmals in der Grundgesamtheit und desto näher liegt der Stichprobenmittelwert $\bar{x}$ am Parameter der Grundgesamtheit μ.

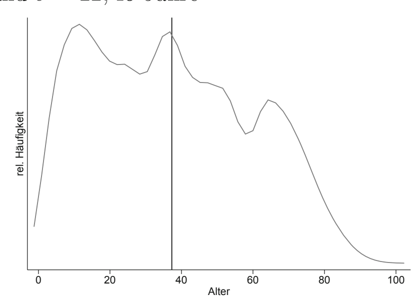
Die *Kennwerteverteilung* ist eine Wahrscheinlichkeitsverteilung. Sie gibt die Wahrscheinlichkeit von Kennwerten, z. B. des arithmetischen Mittels, in Stichproben gleichen Umfangs n an. Aus dem Zentralen Grenzwertsatz folgt – wie wir gesehen haben –, dass sich Mittelwerte in Stichproben bei hinreichend großem n normal um den Mittelwert der Grundgesamtheit μ

verteilen. Das arithmetische Mittel (der Erwartungswert) der Stichprobenmittelwerteverteilung ist μ. Die Standardabweichung einer Kennwerteverteilung wird als *Standardfehler* bzw. Standardschätzfehler bezeichnet. Der Standardfehler gibt an, wie weit Kennwerte in Stichproben vom Parameter der Grundgesamtheit abweichen. Der Standardfehler des Mittelwerts ist $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Mit Hilfe der Kennwerteverteilung lässt sich die Abweichung der Kennwerte in Stichproben vom Parameter der Grundgesamtheit angeben.

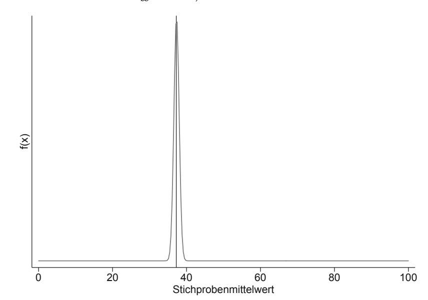
In Abbildung 10.7 sind die drei Verteilungen dargestellt. Oben findet sich die Altersverteilung der bundesdeutschen Bevölkerung im Jahr 1974. Auf der x-Achse ist das Merkmal Alter abgetragen. Ausnahmsweise sind hier die Parameter der Grundgesamtheit bekannt. Der Altersdurchschnitt in der Grundgesamtheit ($\mu = 37,27$ Jahre) ist durch einen senkrechten Strich in der Abbildung eingezeichnet. Darunter ist die Verteilung der arithmetischen Mittel des Alters in Stichproben vom Umfang $n = 1000$ abgebildet. Auf der x-Achse sind Stichprobenmittelwerte $\bar{x}_i$ – hier Altersdurchschnitte – abgetragen. Die Stichprobenmittelwerte verteilen sich mit einem Standardfehler von 0,71 normal um den Mittelwert der Grundgesamtheit ($\mu = 37,27$ Jahre). In der unteren Abbildung ist die Altersverteilung in einer *einzigen* Stichprobe vom Umfang $n = 1000$ angegeben. Der Alterdurchschnitt liegt in dieser Stichprobe bei $\bar{x} = 36{,}4$ Jahren, die Standardabweichung bei $s = 22{,}41$ Jahren. In dieser Stichprobe ist der Altersdurchschnitt niedriger als in der Grundgesamtheit.

Abbildung 10.7: Grundgesamtheit, Kennwerteverteilung und Stichprobe

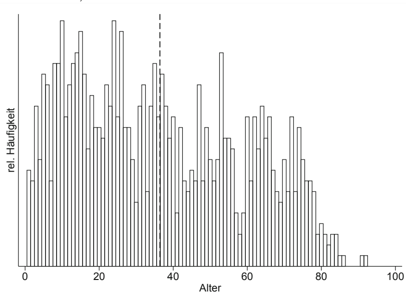
Altersverteilung der bundesdeutschen Bevölkerung 1974.
$\mu = 37,27$ Jahre und $\sigma = 22,46$ Jahre



Stichprobenmittelwerteverteilung für $n = 1000$
$E(\bar{X}) = \mu = 37,27$ Jahre und $\sigma_{\bar{x}} = 0,71$ Jahre



Altersverteilung in *einer* Stichprobe mit $n = 1000$
$\bar{x} = 36,4$ Jahre und $s = 22,41$ Jahre



Aufgaben zu Wahrscheinlichkeitsverteilungen

1. Wie viel Prozent der Fläche (der Werte) liegen a) links und b) rechts von folgenden z-Werten: $z = -2{,}78$; $z = -0{,}1$; $z = 0{,}9$; $z = 1{,}96$?
2. Bitte bestimmen Sie, wie viel Prozent der Fläche bei der Standardnormalverteilung zwischen $z = -2$ und $z = 2$ liegen.
3. Wodurch werden verschiedene Normalverteilungen charakterisiert, wodurch unterscheiden sich diese?
4. Gegeben ist eine Normalverteilung $N(x|20; 16)$ mit einem Mittelwert von 20 und einer Varianz von 16. Bitte berechnen Sie, wie viel Prozent der Fläche in das Intervall zwischen x=20 und x=23 fällt.
5. Was besagt der „Zentrale Grenzwertsatz"?
6. Gegeben ist die Altersverteilung der Bevölkerung der BRD. Die Stichprobenmittelwerte aus dieser Altersverteilung sind nach dem zentralen Grenzwertsatz normalverteilt mit einem Mittelwert von 37,9 Jahren und einem Standardfehler (=Standardabweichung der Stichprobenmittelwerte) von $\sigma_{\bar{x}} = 0{,}7$; d. h. $N(\bar{x}|37{,}9; 0{,}7^2)$. In wieviel Prozent aller Stichproben erhalten Sie einen Altersdurchschnitt zwischen 36,9 und 38,9 Jahren?
7. Antworten Sie mit richtig oder falsch.
   Der Standardfehler eines Stichprobenkennwertes beschreibt
   a) die Standardabweichung in der Grundgesamtheit,
   b) den Fehler, der durch Messungenauigkeiten entsteht,
   c) die Standardabweichung in der Stichprobe,
   d) die Standardabweichung der Kennwerteverteilung.

11 Konfidenzintervalle

Im letzten Kapitel haben wir uns mit der Abweichung der Kennwerte in Stichproben vom Parameter der Grundgesamtheit beschäftigt. In der Praxis haben wir nur Daten einer einzigen Stichprobe. Mit diesen Daten möchten wir Aussagen über einen Parameter der Grundgesamtheit treffen. Ein typisches Beispiel sind Wahlprognosen: Dort werden per Umfrage die Stimmenanteile für die einzelnen Parteien bei einem Teil der Wähler ermittelt. Wissen möchte man natürlich, wie die einzelnen Parteien bei allen Wählern abschneiden.

Die Schätzung der Populationsparameter kann als *Punkt-* oder als *Intervallschätzung* vorgenommen werden. Bei einer Punktschätzung wird der Parameter der Grundgesamtheit durch einen einzigen Wert der Stichprobe geschätzt. In der oben dargestellten Wahlumfrage (vgl. Tabelle 9.1 auf Seite 194) könnten z. B. die 42,5 % der gültigen Stimmen, die die CDU unter den Befragten erhielt, als Schätzwert für den Prozentsatz der gültigen Stimmen bei allen Wählern verwendet werden. Weil Punktschätzer mit der zufälligen Zusammensetzung der Stichprobe variieren, gibt man Bereiche an, in denen die Parameter der Grundgesamtheit mit einer gewissen Wahrscheinlichkeit liegen.

11.1 Punktschätzung

Als Punktschätzer des arithmetischen Mittels in der Grundgesamtheit μ wird das in einer konkreten Stichprobe beobachtete arithmetische Mittel $\bar{x}$ verwandt:

$$\hat{\mu} = \bar{x} \,. \tag{11.1}$$

$\bar{x}$ ist eine Realisation der Zufallsvariablen $\bar{X}$ – arithmetische Mittelwerte in Stichproben. $\bar{X}$ wird als Schätzfunktion (Schätzer, *estimator*) für μ

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_11,

benutzt. Im ALLBUS 2004 wurde für männliche Befragte in Westdeutschland eine durchschnittliche Körpergröße von $\bar{x} = 178$ cm ermittelt, die wir zur Schätzung von μ verwenden können: $\hat{\mu} = 178$ cm.

Zur Schätzung eines Anteils der Grundgesamtheit verwenden wir den Anteil in einer konkreten Stichprobe p.

$$\hat{\theta} = p \tag{11.2}$$

Die Schätzung des Stimmenanteils der CDU θ durch den Stimmenanteil p der Stichprobe ist: $\hat{\theta} = 0{,}425 = 42{,}5\%$. Die Schätzfunktion ist die Zufallsvariable P – Anteile in Stichproben.

Schätzer haben drei wünschenswerte Eigenschaften:

- Erwartungstreue,
- Effizienz und
- Konsistenz.

Eine Schätzfunktion ist *erwartungstreu*, wenn der Erwartungswert der Funktion, d. h. deren Mittelwert, dem zu schätzenden Parameter der Grundgesamtheit entspricht. Der Erwartungswert der Verteilung der arithmetischen Mittel in Stichproben $\bar{X}$ ist μ. $\bar{X}$ ist daher ein unverzerrter Schätzer von μ. Dagegen ist die Varianz in Stichproben kein unverzerrter Schätzer der Varianz in der Grundgesamtheit. Die Varianz in einer Stichprobe unterschätzt die Varianz in der Grundgesamtheit. Aus diesem Grund wird die Varianz einer Stichprobe s^2 mit einem Korrekturfaktor multipliziert, wenn sie zur Schätzung der Varianz in der Grundgesamtheit σ^2 verwandt wird (vgl. Gleichung 11.6, S. 263). Ein verzerrter Schätzer unter- oder überschätzt den Parameter der Grundgesamtheit im Durchschnitt.

Ein Schätzer ist *effizient*, wenn er einen kleineren Standardfehler hat als andere Schätzer. Als Beispiel soll die Körpergröße herangezogen werden. Die Körpergröße ist ein normalverteiltes Merkmal, bei dem arithmetisches Mittel und Median identisch sind. Zur Schätzung der zentralen Lage der Körpergröße der Grundgesamtheit könnte man nun den Median oder das arithmetische Mittel der Stichprobe heranziehen. Es lässt sich zeigen, dass

der Median in Stichproben $\tilde{X}$ einen größeren Standardfehler aufweist als das arithmetische Mittel in Stichproben $\bar{X}$.

Ein Schätzer wird als *konsistent* bezeichnet, wenn mit zunehmendem Stichprobenumfang die Wahrscheinlichkeit eines Abstands zum zu schätzenden Parameter geringer wird. Konsistenz kennzeichnet das Verhalten eines Schätzers bei Vergrößerung der Stichprobe.

11.2 Konfidenzintervall für den Mittelwert μ

Bei einem Konfidenzintervall wird ein Bereich angegeben, indem der gesuchte Paramter der Grundgesamtheit vermutet wird.

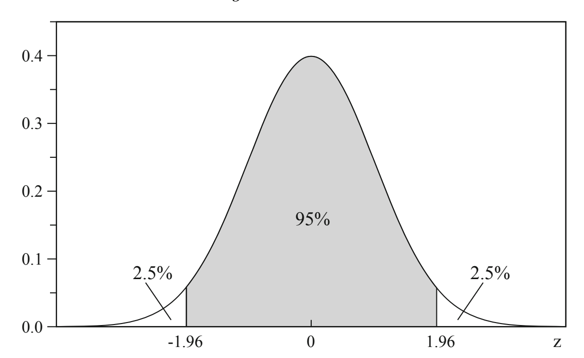
Wahrscheinlichkeitsintervalle für Stichprobenmittelwerte

Durch das Zentrale Grenzwerttheorem wissen wir, dass sich Stichprobenmittelwerte normalverteilen, wenn die gezogenen Stichproben hinreichend groß sind. Deshalb können wir – bei Kenntnis der Varianz σ^2 und des Mittelwertes μ der Grundgesamtheit – berechnen, wieviel Prozent der Stichprobenmittelwerte $\bar{X}$ in bestimmten Grenzen liegen (vgl. S. 242) und umgekehrt, in welchen Grenzen sich ein bestimmter Prozentsatz der Stichprobenmittelwerte befindet. Dieser Prozentsatz gibt auch die Wahrscheinlichkeit an, mit der *ein* Stichprobenmittelwert in diesem Intervall erwartet werden darf. Solche *Bereiche, in denen die Stichprobenmittelwerte mit einer gewissen Wahrscheinlichkeit liegen*, werden als *Wahrscheinlichkeitsintervalle* bezeichnet.

Gesucht seien z. B. die *Grenzen* des Intervalls, innerhalb dessen sich 95 % der Werte einer Standardnormalverteilung *symmetrisch* zum Mittelwert befinden. Wir bestimmen also jetzt nicht die Fläche aufgrund vorgegebener Grenzen, sondern die Grenzen anhand einer vorgegebenen Fläche – nämlich 95 %. Das gesuchte Intervall, das in Abbildung 11.1 durch die schraffierte Fläche repräsentiert wird, umfasst 95 % der Werte, die beiden nicht schraffierten Flächen rechts und links beinhalten jeweils 2,5 % der Fläche. Links vom unteren Grenzwert befinden sich also 2,5 % der Fläche, links vom oberen Grenzwert dagegen 97,5 %. Dementsprechend befinden sich die beiden Grenzwerte an den Stellen $z_{0,025}$ bzw. $z_{0,975}$. Man sucht also innerhalb der z-Tabelle in Anhang A die Fläche 0,975 bzw. 0,025 und liest dann am Rand der Tabelle den zugehörigen z-Wert ab. Wie man der

z-Tabelle entnehmen kann, entspricht der unteren Fläche ein Wert von $z_{0,025} = -1{,}96$ und der oberen ein Wert von $z_{0,975} = 1{,}96$. 95 % der Fläche befinden sich bei einer Standardnormalverteilung also zwischen $-1{,}96$ und $+1{,}96$.

Abbildung 11.1: 95 %-Wahrscheinlichkeitsintervall einer *Standardnormalverteilung*

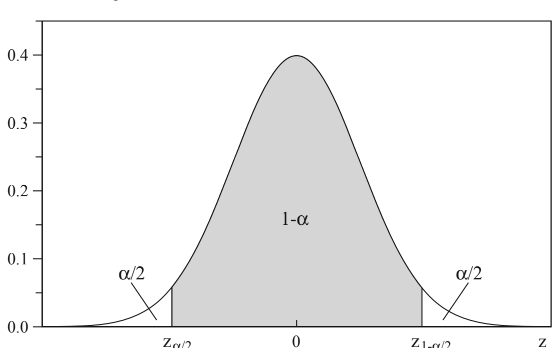


Allgemein bezeichnet man die Wahrscheinlichkeit, dass ein Wert *nicht* in das Wahrscheinlichkeitsintervall fällt, mit α. Die Wahrscheinlichkeit dafür, dass ein Wert in das Wahrscheinlichkeitsintervall fällt, wird mit $1 - \alpha$ bezeichnet. Bei Intervallen, die symmetrisch zum Mittelwert μ liegen, befinden sich am linken und rechten Rand der Verteilung $\frac{\alpha}{2}$ der Gesamtfläche. Die Grenzen liegen also bei der Standardnormalverteilung an den Stellen $z_{\frac{\alpha}{2}}$ und $z_{1-\frac{\alpha}{2}}$, wie man in Abbildung 11.2 sehen kann. Solche zum Mittelpunkt symmetrischen Wahrscheinlichkeitsintervalle nennt man **zweiseitige Intervalle**.

Handelt es sich nicht um eine standardisierte, sondern um eine beliebige Normalverteilung – und um eine solche handelt es sich ja auch bei der Stichprobenmittelwerteverteilung – dann müssen die Grenzen des Intervalls $z_{\frac{\alpha}{2}}$ und $z_{1-\frac{\alpha}{2}}$ destandardisiert werden, indem die z-Transformation (vgl. Gleichung 10.21 auf Seite 247) rückgängig gemacht wird.

Setzt man $z_{\frac{\alpha}{2}}$ und $z_{1-\frac{\alpha}{2}}$ für z in die Gleichung $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ ein und löst beide Gleichungen nach $\bar{x}$ auf, dann erhält man als untere Grenze $\mu + z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$ und als obere Grenze $\mu + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$.

Abbildung 11.2: Wahrscheinlichkeitsintervall einer *Standardnormalverteilung*

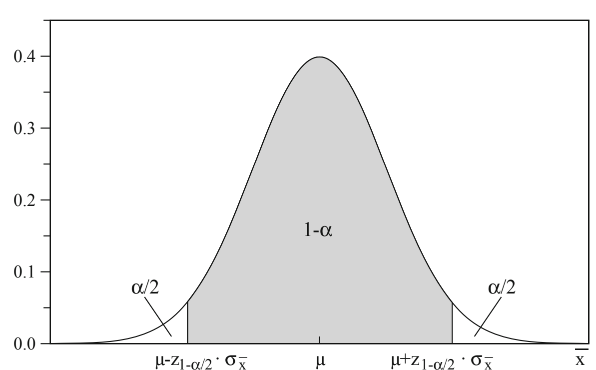


Aufgrund der Symmetrie der Verteilung ist $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$, weshalb man für die untere Grenze auch $\mu - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$ schreiben kann. Die Formel zur Berechnung des **Wahrscheinlichkeitsintervalls** einer Stichprobenmittelwerteverteilung lautet daher (vgl. Abbildung 11.3):

$$\mu - z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} \quad \leq \bar{X} \leq \quad \mu + z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}\,. \tag{11.3}$$

Die Berechnung des Wahrscheinlichkeitsintervalls soll wiederum am Beispiel der Altersverteilung der bundesdeutschen Bevölkerung verdeutlicht werden. Das Durchschnittsalter der Bundesbürger betrug 1974 $\mu = 37{,}27$ Jahre, die Standardabweichung $\sigma = 22{,}46$ Jahre. Der Standardfehler des Mittelwerts $\sigma_{\bar{x}}$ gibt die Breite der Kennwerteverteilung an und wird durch $\frac{\sigma}{\sqrt{n}}$ berechnet. Wir möchten nun wissen, in welchem Intervall sich 95 % der arithmetischen Mittel (Altersdurchschnitte) in Stichproben vom Umfang $n = 1000$ aus dieser Grundgesamtheit befinden. α beträgt also 5 % bzw. 0,05, $1 - \alpha$ somit 95 % bzw. 0,95.

Abbildung 11.3: Wahrscheinlichkeitsintervall einer *Stichprobenmittelwerteverteilung*



$$
\begin{aligned}
37{,}27 - z_{1-\frac{0{,}05}{2}} \cdot \frac{22{,}46}{\sqrt{1000}} &\leq \bar{X} \leq 37{,}27 + z_{1-\frac{0{,}05}{2}} \cdot \frac{22{,}46}{\sqrt{1000}} \\
37{,}27 - z_{0{,}975} \cdot \frac{22{,}46}{\sqrt{1000}} &\leq \bar{X} \leq 37{,}27 + z_{0{,}975} \cdot \frac{22{,}46}{\sqrt{1000}} \\
37{,}27 - 1{,}96 \cdot 0{,}71 &\leq \bar{X} \leq 37{,}27 + 1{,}96 \cdot 0{,}71 \\
35{,}88 &\leq \bar{X} \leq 38{,}66
\end{aligned}
$$

In 95 % der möglichen Stichproben vom Umfang $n = 1000$ liegt der Altersdurchschnitt zwischen 35,88 und 38,66 Jahren.

Konfidenzintervall für den Mittelwert μ bei bekannter Varianz der Grundgesamtheit

Ebenso wie wir ein Intervall um μ gelegt haben, können wir nun ein Intervall um einen Stichprobenmittelwert $\bar{x}$ legen. Liegt ein Stichprobenmittelwert im grau schraffierten Bereich in Abbildung 11.3, dann ist μ nicht weiter als $\pm z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$ von diesem Stichprobenmittelwert $\bar{x}$ entfernt.

Wenn nämlich $(1-\alpha)\,\%$ der möglichen Stichprobenmittelwerte nicht weiter als $\pm z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$ (Intervall) vom Mittelwert der Grundgesamtheit μ entfernt sind, dann ist auch der Mittelwert der Grundgesamtheit μ nicht weiter als $\pm z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$ vom Stichprobenmittelwert $\bar{x}$ bei $(1-\alpha)\,\%$ der möglichen Stichproben entfernt. Konkret: Weichen 95 % der Stichprobenmittelwerte nicht weiter als $\pm 1{,}96 \cdot \sigma_{\bar{x}}$ vom Mittelwert der Grundgesamtheit μ ab, dann ist μ bei 95 % der Stichprobenmittelwerte auch nicht weiter als $\pm 1{,}96 \cdot \sigma_{\bar{x}}$ von $\bar{x}$ entfernt. In 5 % der Stichproben ist μ weiter entfernt. Dieser Prozentsatz kann als Wahrscheinlichkeit interpretiert werden: Die Wahrscheinlichkeit dafür, dass der Mittelwert der Grundgesamtheit μ im Intervall $\pm 1{,}96 \cdot \sigma_{\bar{x}}$ um einen möglichen Stichprobenmittelwert $\bar{x}$ liegt, beträgt 95 %, die Wahrscheinlichkeit, dass μ außerhalb dieses Intervalls liegt, beträgt lediglich 5 %.

Berechnung von Konfidenzintervallen

Solche Bereiche, in denen ein unbekannter Parameter der Grundgesamtheit vermutet wird, werden als Vertrauens- oder **Konfidenzintervalle** bezeichnet. Die Bildung von Konfidenzintervallen erfolgt im Prinzip ebenso wie die von Wahrscheinlichkeitsintervallen. Die untere Grenze $-z_{1-\frac{\alpha}{2}}$ und die obere Grenze $z_{1-\frac{\alpha}{2}}$ müssen wiederum für z in die Gleichung $z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}}$ eingesetzt werden. Da wir hier den Wert μ suchen, müssen beide Gleichungen jedoch nach μ aufgelöst werden. Als Konfidenzintervall erhalten wir

$$\underbrace{\bar{x} - z_{(1-\frac{\alpha}{2})} \cdot \sigma_{\bar{x}}}_{\text{untere Grenze}} \quad \leq \mu \leq \quad \underbrace{\bar{x} + z_{(1-\frac{\alpha}{2})} \cdot \sigma_{\bar{x}}}_{\text{obere Grenze}} \;. \tag{11.4}$$

Setzt man $\frac{\sigma}{\sqrt{n}}$ für den Standardfehler des Mittelwertes $\sigma_{\bar{x}}$ ein, ergibt sich folgende Gleichung:

$$\underbrace{\bar{x} - z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{untere Grenze}} \quad \leq \mu \leq \quad \underbrace{\bar{x} + z_{(1-\frac{\alpha}{2})} \cdot \frac{\sigma}{\sqrt{n}}}_{\text{obere Grenze}} \;. \tag{11.5}$$

Die Berechnung des Konfidenzintervalls kann wiederum am Beispiel des Durchschnittsalters der bundesdeutschen Bevölkerung im Jahr 1974 nachvollzogen werden. Wir wählen (mit Hilfe des Programmes ALTMIHI aus

GSTAT) zufällig eine Stichprobe mit 1.000 Befragten aus und ermitteln für diese Stichprobe einen Altersdurchschnitt $\bar{x}$ von 38,11 Jahren. Die Standardabweichung der Grundgesamtheit σ beträgt 22,46 Jahre. Die Intervallgrenzen können nach Formel 11.5 bestimmt werden:

$$\begin{aligned}
38{,}11 - z_{(1-\frac{0{,}05}{2})} \cdot \frac{22{,}46}{\sqrt{1000}} &\leq \mu \leq 38{,}11 + z_{(1-\frac{0{,}05}{2})} \cdot \frac{22{,}46}{\sqrt{1000}} \\
38{,}11 - 1{,}96 \cdot 0{,}71 &\leq \mu \leq 38{,}11 + 1{,}96 \cdot 0{,}71 \\
36{,}72 &\leq \mu \leq 39{,}50\,.
\end{aligned}$$

Konfidenzintervalle werden häufig in eckigen Klammern angegeben: $[36{,}72; 39{,}5]$. Mit 95%iger Wahrscheinlichkeit enthält das Intervall $[36{,}72; 39{,}5]$ den Altersdurchschnitt der Population.

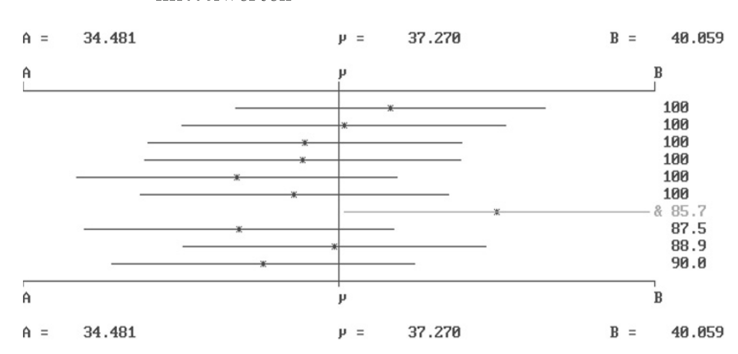
Interpretation von Konfidenzintervallen

Da der Altersdurchschnitt einer Stichprobe $\bar{x}$ von deren zufälliger Zusammensetzung abhängt, werden wir für unterschiedliche Stichproben unterschiedliche Altersdurchschnitte und damit unterschiedliche Konfidenzintervalle für μ erhalten. Mit dem Programm SIMKONOR aus GSTAT sind zehn verschiedene Stichproben mit jeweils 1.000 Personen aus der Altersverteilung der Bundesdeutschen gezogen worden. Für jede der Stichproben wurde das 95%ige Konfidenzintervall berechnet. Die Stichprobenmittelwerte sind in Abbildung 11.4 durch Sternchen gekennzeichnet, die Konfidenzintervalle durch Linien. Weil die Stichprobenmittelwerte unterschiedlich sind, unterscheidet sich auch die Lage der Konfidenzintervalle.

Da wir hier ausnahmsweise den Mittelwert der Grundgesamtheit μ kennen, können wir angeben, ob ein Konfidenzintervall μ umschließt oder nicht. In neun der zehn Stichproben liegt μ, das in der Graphik als senkrechter Strich eingezeichnet wurde, tatsächlich im berechneten Konfidenzintervall. In der siebten Stichprobe liegt der Altersdurchschnitt der Grundgesamtheit, nämlich 37,27 Jahre, außerhalb des Konfidenzintervalls.

Normalerweise kennt man den Wert von μ nicht, weshalb wir nicht angeben können, ob ein konkretes Konfidenzintervall μ tatsächlich einschließt oder nicht. In 95 % der möglichen Stichproben aus einer Grundgesamtheit

Abbildung 11.4: Konfidenzintervalle bei unterschiedlichen Stichprobenmittelwerten



Ausgabe des Programms SIMKONOR aus GSTAT

wird das Konfidenzintervall den Parameter der Grundgesamtheit enthalten. Wenn wir – wie im Beispiel – wiederholt Stichproben vom gleichen Umfang ziehen, so werden langfristig, d. h. bei einer großen Zahl von Stichproben, etwa 95 % der Konfidenzintervalle den Mittelwert der Grundgesamtheit beinhalten. In etwa 5 % der Stichproben wird das um den Stichprobenmittelwert gelegte Konfidenzintervall den Parameter μ nicht einschließen, und wir irren uns bei der Schätzung – wie hier in der siebten Stichprobe. Aus diesem Grund wird α auch als **Irrtumswahrscheinlichkeit** und $1 - \alpha$ als **Vertrauenswahrscheinlichkeit** bezeichnet.

Erhöhung der Vertrauenswahrscheinlichkeit

Ist uns dieser Schluss zu unsicher, dann können wir die Vertrauenswahrscheinlichkeit z. B. auf $1 - \alpha = 0{,}99$, also 99 %, erhöhen. Die Irrtumswahrscheinlichkeit beträgt $\alpha = 0,01$. Die $z-$Werte zu den Quantilen $0,05$ und $0,995$ sind $\pm 2{,}58$.

$$38{,}11 - z_{(1-\frac{0{,}01}{2})} \cdot \frac{22{,}46}{\sqrt{1000}} \leq \mu \leq 38{,}11 + z_{(1-\frac{0{,}01}{2})} \cdot \frac{22{,}46}{\sqrt{1000}}$$

$$38{,}11 - 2{,}58 \cdot 0{,}71 \leq \mu \leq 38{,}11 + 2{,}58 \cdot 0{,}71$$

$$36{,}28 \leq \mu \leq 39{,}94$$

Das Intervall [36,28;39,94] beinhaltet mit 99%iger Wahrscheinlichkeit das Durchschnittsalter der bundesdeutschen Bevölkerung.

Mit zunehmender Vertrauenswahrscheinlichkeit wird das Konfidenzintervall *breiter*. Die höhere Sicherheit beim Schließen geht also mit einer ungenaueren Schätzung des unbekannten Mittelwerts einher. Der Extremfall, dass wir unseren Schluss mit 100%iger Sicherheit tätigen wollten, würde die Intervallgrenzen auf $-\infty$ bzw. $+\infty$ ausdehnen. Die dazugehörende Aussage „Mit 100%iger Wahrscheinlichkeit überdeckt das Intervall von $-\infty$ bis $+\infty$ den Parameter μ“ ist allerdings nicht informativ.

Konfidenzintervall für den Mittelwert μ bei unbekannter Varianz der Grundgesamtheit

Im vorangegangenen Beispiel sind wir von einer bekannten Varianz σ^2 und Standardabweichung σ der Grundgesamtheit ausgegangen. Normalerweise ist σ jedoch nicht bekannt, und damit kann auch der Standardfehler des Mittelwerts $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ nicht berechnet werden. Als Schätzwert der Standardabweichung der Grundgesamtheit σ verwendet man dann die *Standardabweichung der Stichprobe* s. Die Varianz in der Stichprobe s^2 ist allerdings kein erwartungstreuer Schätzer der Varianz in der Grundgesamtheit. Daher muss s^2 mit dem Faktor $n/(n-1)$ korrigiert werden, um den Schätzwert für die Varianz der Grundgesamtheit $\hat{\sigma}^2$ zu erhalten.

$$\hat{\sigma}^2 = s^2 \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \cdot \frac{n}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n-1}} \tag{11.6}$$

Die Schätzung der Varianz der Grundgesamtheit auf Basis der Stichprobe $\hat{\sigma}^2$ unterscheidet sich von der Stichprobenvarianz s^2 also nur durch den

Nenner. Im ersteren Fall wird durch $n - 1$, im letzteren jedoch durch n dividiert. Bei großen Stichproben ist der Berechnungsunterschied bedeutungslos. Die geschätzte Standardabweichung der Grundgesamtheit $\hat{\sigma}$ ist $\sqrt{\hat{\sigma}^2}$.

Durch die Schätzung von σ durch die Stichprobendaten nehmen wir einen zusätzlichen Unsicherheitsfaktor in Kauf, da die Standardabweichung in der Stichprobe nicht identisch mit der Standardabweichung der Grundgesamtheit sein muss. Dieser Tatsache trägt man Rechnung, indem man zur Bestimmung der Konfidenzintervalle nicht die z-, sondern die breitere *t-Verteilung* heranzieht.

t-Verteilung

Die t-Verteilung ähnelt der Normalverteilung (vgl. Abbildung 11.5), variiert aber mit der Zahl der „Freiheitsgrade" (abgekürzt df = degrees of freedom). t-Verteilungen sind flacher und breiter als Normalverteilungen, weisen aber denselben glockenförmigen Verlauf auf. Mit zunehmender Zahl an Freiheitsgraden nähert sich die t-Verteilung der Normalverteilung an, wie man in Abbildung 11.5 sieht. Die t-Verteilung ist symmetrisch zum Mittelwert null. Die Varianz der t-Verteilung $(df/(df-2))$ sinkt mit zunehmender Zahl der Freiheitsgrade.[1] Bereits bei 29 Freiheitsgraden unterscheidet sich die t-Verteilung kaum noch von der z-Verteilung. Die Freiheitsgrade lassen sich als $df = n - 1$ bestimmen, wobei n der Stichprobenumfang ist. Auch die t-Verteilung liegt in Tabellen vor (vgl. Anhang A).

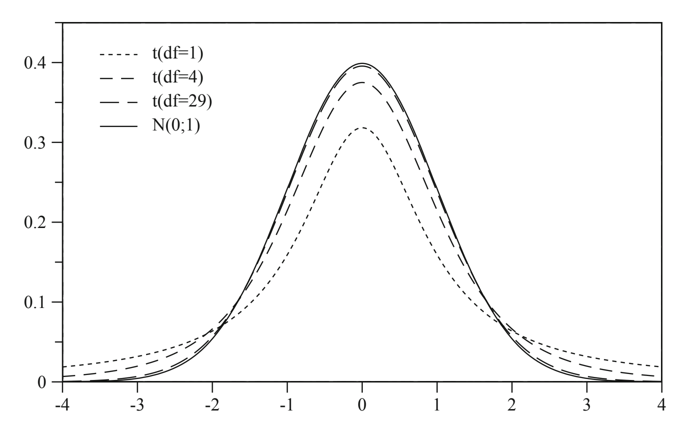
Für das Konfidenzintervall erhält man

$$\underbrace{\bar{x} - t_{(1-\frac{\alpha}{2};n-1)} \cdot \frac{\hat{\sigma}}{\sqrt{n}}}_{\text{untere Grenze}} \quad \leq \mu \leq \quad \underbrace{\bar{x} + t_{(1-\frac{\alpha}{2};n-1)} \cdot \frac{\hat{\sigma}}{\sqrt{n}}}_{\text{obere Grenze}} \ . \tag{11.7}$$

Mit Hilfe des Programms `ALTER` aus `GSTAT` haben wir eine Stichprobe von 81 Personen aus der bundesdeutschen Bevölkerung gezogen. In der Stichprobe beträgt das Durchschnittsalter $\bar{x}$ = 38,57 Jahre und die Varianz s^2 = 423,1249. Nach Gleichung 11.6 von Seite 263 schätzen wir die Standardabweichung des Alters in der Grundgesamtheit auf

1 Für $df < 2$ ist die Varianz der Verteilung nicht definiert.

Abbildung 11.5: t-Verteilungen in Abhängigkeit vom Freiheitsgrad



$\hat{\sigma} = \sqrt{423{,}1249 \cdot \frac{81}{80}} = 20{,}7$ Jahre. Wir möchten nun mit 95%iger Vertrauenswahrscheinlichkeit das Konfidenzintervall des Altersdurchschnitts der Gesamtbevölkerung berechnen; α ist also 0,05 bzw 5 %.

$$\begin{aligned}
38{,}57 - t_{(1-\frac{0{,}05}{2};81-1)} \cdot \frac{20{,}7}{\sqrt{81}} &\leq \mu \leq 38{,}57 + t_{(1-\frac{0{,}05}{2};81-1)} \cdot \frac{20{,}7}{\sqrt{81}} \\
38{,}57 - t_{(0{,}975;80)} \cdot \frac{20{,}7}{\sqrt{81}} &\leq \mu \leq 38{,}57 + t_{(0{,}975;80)} \cdot \frac{20{,}7}{\sqrt{81}} \\
38{,}57 - 1{,}990 \cdot 2{,}3 &\leq \mu \leq 38{,}57 + 1{,}990 \cdot 2{,}3 \\
33{,}99 &\leq \mu \leq 43{,}15
\end{aligned}$$

Mit 95%iger Sicherheit überdeckt das Intervall [34 Jahre; 43 Jahre] den Altersdurchschnitt in der Grundgesamtheit. Weil die Stichprobe so klein ist, ist der Standardfehler des Mittelwertes groß. Das Konfidenzintervall fällt daher sehr breit aus.

Bei einer Stichprobengröße von 121 Befragten und damit 120 Freiheitsgraden, beträgt der t-Wert für ein zweiseitiges Konfidenzintervall bei einer Vertrauenswahrscheinlichkeit von 95 % $t_{(1-\frac{\alpha}{2});120} = 1{,}98$, während der

zu dieser Vertrauenswahrscheinlichkeit gehörende z-Wert $z_{1-\frac{\alpha}{2}} = 1{,}96$ ist. Bei einer Stichprobe mit 1.000 Befragten fällt die Differenz zwischen t- und z-Wert bereits nicht mehr ins Gewicht.

t-Tabellen beinhalten t-Verteilungen bis zu 200 Freiheitsgraden. Bei großem Stichprobenumfang und großer Zahl an Freiheitsgraden, kann die z-Verteilung verwendet werden.

$$\underbrace{\bar{x} - z_{(1-\frac{\alpha}{2})} \cdot \frac{\hat{\sigma}}{\sqrt{n}}}_{\text{untere Grenze}} \leq \mu \leq \underbrace{\bar{x} + z_{(1-\frac{\alpha}{2})} \cdot \frac{\hat{\sigma}}{\sqrt{n}}}_{\text{obere Grenze}} \tag{11.8}$$

Ein abschließendes Beispiel: Im ALLBUS 2004 betrug das Körpergewicht der in Westdeutschland befragten 963 Frauen durchschnittlich $\bar{x} = 69$ kg und $\hat{\sigma} = 14,2$ kg. (Das Minimum der Verteilung liegt bei 37 kg, das Maximum bei 160 kg.) Da der Stichprobenumfang relativ groß ist, wenden wir die z-Tabelle zur Berechnung eines 99%igen Konfidenzintervalls an.

$$\begin{aligned}
69 - z_{(1-\frac{0,01}{2})} \cdot \frac{14{,}2}{\sqrt{963}} &\leq \mu \leq 69 + z_{(1-\frac{0,01}{2})} \cdot \frac{14{,}2}{\sqrt{963}} \\
69 - 2{,}58 \cdot \frac{14{,}2}{\sqrt{963}} &\leq \mu \leq 69 + 2{,}58 \cdot \frac{14{,}2}{\sqrt{963}} \\
67{,}8 &\leq \mu \leq 70{,}2
\end{aligned}$$

Das Körpergewicht westdeutscher Frauen liegt mit 99%iger Wahrscheinlichkeit im Intervall von 67,8 bis 70,2 kg.

11.3 Konfidenzintervall für den Anteilswert θ

Vor der Bundestagswahl 1994 ermittelte die Forschungsgruppe Wahlen in einer Umfrage einen Stimmenanteil von 7% für die FDP. Insgesamt wurden 1.250 Personen befragt (vgl. Tabelle 9.1 auf Seite 194). Wir möchten natürlich wissen, wie die FDP bei allen Wählern abschneidet. Wir suchen also den unbekannten Anteil bzw. Prozentwert θ (theta) der FDP in der Grundgesamtheit.

Die Logik bei der Bildung eines Konfidenzintervalls für Anteilswerte entspricht der für Mittelwerte. In Kapitel 10 wurde gezeigt, dass sich auch Häufigkeiten und Anteilswerte bei hinreichend großem Stichprobenumfang n normal um den Parameter der Grundgesamtheit θ mit einem Standardfehler σ_p verteilen. Als eine Faustregel für ein „genügend großes n" gilt $n \cdot \theta \cdot (1 - \theta) \geq 9$. Weil θ unbekannt ist, wird der Stichprobenanteil p herangezogen: $n \cdot p \cdot (1 - p) \geq 9$. Im Beispiel erhalten wir also $1250 \cdot 0{,}07 \cdot 0{,}93 = 81{,}375$. Auch die Bedingungen in Gleichung 10.24 (S. 249) sind erfüllt, wie man leicht nachrechnen kann.

Die z-Transformation für die *Anteilswerteverteilung* lautet

$$z = \frac{p - \theta}{\sigma_p} \,. \tag{11.9}$$

Das *Konfidenzintervall für den unbekannten Anteilswert θ* der Grundgesamtheit wird nun gebildet, indem $-z_{1-\frac{\alpha}{2}}$ als untere Grenze und $z_{1-\frac{\alpha}{2}}$ als obere Grenze in Gleichung 11.9 eingesetzt und beide Gleichungen nach θ aufgelöst werden. Das Konfidenzintervall für den Anteil der Grundgesamtheit θ berechnet sich demnach nach

$$\underbrace{p - z_{(1-\frac{\alpha}{2})} \cdot \sigma_p}_{\text{untere Grenze}} \quad \leq \theta \leq \quad \underbrace{p + z_{(1-\frac{\alpha}{2})} \cdot \sigma_p}_{\text{obere Grenze}} \,. \tag{11.10}$$

Der **Standardfehler des Stichprobenanteilswertes** σ_p wird nach Gleichung 10.7 (S. 10.7)

$$\sigma_p = \sqrt{\frac{\theta \cdot (1 - \theta)}{n}}$$

ermittelt. Weil θ nicht bekannt ist, schätzen wir σ_p durch den Anteilswert in der Stichprobe:

$$\hat{\sigma}_p = \sqrt{\frac{p \cdot (1 - p)}{n}} \,. \tag{11.11}$$

Setzen wir $\hat{\sigma}_p$ für σ_p in Gleichung (11.10) ein, so erhalten wir

$$\underbrace{p - z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}}_{\text{untere Grenze}} \leq \theta \leq \underbrace{p + z_{(1-\frac{\alpha}{2})} \cdot \sqrt{\frac{p \cdot (1-p)}{n}}}_{\text{obere Grenze}} . \qquad (11.12)$$

Anhand dieser Formel können wir nun das 95%ige Konfidenzintervall für den Stimmenanteil der FDP ermitteln.

$$\begin{aligned} 0{,}07 - z_{(1-\frac{0{,}05}{2})} \cdot \sqrt{\frac{0{,}07 \cdot 0{,}93}{1250}} &\leq \theta \leq 0{,}07 + z_{(1-\frac{0{,}05}{2})} \cdot \sqrt{\frac{0{,}07 \cdot 0{,}93}{1250}} \\ 0{,}07 - 1{,}96 \cdot 0{,}0072 &\leq \theta \leq 0{,}07 + 1{,}96 \cdot 0{,}0072 \\ 0{,}0559 &\leq \theta \leq 0{,}0841 \end{aligned}$$

Das Intervall von 5,6 bis 8,4 % der gültigen Stimmen überdeckt den Stimmenanteil der FDP bei allen Wählern mit 95%iger Wahrscheinlichkeit. Die Wahlkampfstrategen der FDP wären mit dieser Schätzung sicher zufrieden, da die Prognose selbst im schlechtesten Fall einen Einzug in den Bundestag beinhaltet.

11.4 Der Einfluss des Stichprobenumfangs

Häufig sind die berechneten Konfidenzintervalle zu breit und damit zu ungenau. Genauere Schätzungen erhält man, wenn man die Vertrauenswahrscheinlichkeit verringert oder den Stichprobenumfang erhöht. Wenn möglich, ist die Erhöhung des Stichprobenumfangs vorzuziehen, da sie nicht mit einem Verlust an Präzision einhergeht. Wie groß der Stichprobenumfang sein muss, um eine bestimmte Genauigkeit der Schätzung zu erzielen, lässt sich relativ einfach bestimmen:

Ziel ist es, die *Konfidenzintervallbreite* (KIB) zu verringern. Die Breite des Konfidenzintervalls ist nichts anderes als der Abstand zwischen der unteren und der oberen Grenze. Für die Stichprobenmittelwerteverteilung beträgt sie

$$KIB = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \sigma_{\bar{x}} \\ = 2 \cdot z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \,. \tag{11.13}$$

Diese Gleichung muss nun nach n aufgelöst werden:

$$n = \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \sigma^2}{KIB^2} \,. \tag{11.14}$$

Ist σ unbekannt, so wird dieses wieder durch $\hat{\sigma}$ geschätzt. Wir können jetzt berechnen, wie groß die Stichprobe der westdeutschen Frauen im ALLBUS 2004 sein müsste, um den unbekannten Mittelwert μ des Körpergewichts bei einer Vertrauenswahrscheinlichkeit von 99 % mit einer Konfidenzintervallbreite von 1 kg zu schätzen. Der zu einem Konfidenzintervall von 99 % gehörende z-Wert beträgt 2,58, $\hat{\sigma} = 14{,}2$ kg.

$$n = \frac{4 \cdot 2{,}58^2 \cdot 14{,}2^2}{1^2} = 5369$$

Um eine Konfidenzintervallbreite von 1 kg zu erhalten, müssten wir also ca. 5.400 Frauen in Westdeutschland befragen. Um die Konfidenzintervallbreite nochmals auf 0,5 kg zu halbieren, müssten wir (bei gleicher Vertrauenswahrscheinlichkeit) 21.475 – also viermal so viele – Personen befragen. Der Stichprobenumfang muss vervierfacht werden, wenn die Konfidenzintervallbreite halbiert werden soll, weil in Gleichung (11.14) durch die quadrierte KIB dividiert wird.

Der Stichprobenumfang für Anteilswerte lässt sich analog herleiten und berechnet sich als

$$n = \frac{4 \cdot z^2_{1-\frac{\alpha}{2}} \cdot \theta(1-\theta)}{KIB^2} \,. \tag{11.15}$$

Ist θ unbekannt, so wird auch hier der Anteilswert p der Stichprobe zur Schätzung herangezogen.

Um vorab den benötigten Stichprobenumfang berechnen zu können, muss man den Standardfehler kennen oder schätzen. Dazu benötigt man entweder σ bzw. θ oder $\hat{\sigma}$ bzw. p. Im obigen Beispiel wurde $\hat{\sigma}$ einer schon durchgeführten Untersuchung verwandt.

Zusammenfassung

In Tabelle 11.4 sind die Punktschätzer und Konfidenzintervalle für Mittel- und Anteilswerte bei unbekannter Varianz der Grundgesamtheit angegeben. In Kapitel 12 wird im Zusammenhang mit einem Testverfahren die Berechnung eines Konfidenzintervalls für die Differenz von zwei Mittelwerten erläutert.

Tabelle 11.1: Punkt- und Intervallschätzung

Parameter	Punkt-schätzer	Standard-fehler*	Konfidenz-intervall
Mittelwert μ	$\bar{x}$	$\frac{\hat{\sigma}}{\sqrt{n}}$	$\bar{x} \pm t_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$
Anteil π	p	$\sqrt{\frac{p(1-p)}{n}}$	$p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p(1-p)}{n}}$

* aus der Stichprobe geschätzt

Die allgemeine Struktur von Konfidenzintervallen lautet

$$\text{Punktschätzer} \pm \text{Quantilwert} \cdot \text{Standardfehler}\,.$$

Konfidenzintervalle lassen sich für alle möglichen Parameter der Grundgesamtheit bestimmen. In Kapitel 8 wurde mit einer linearen Regression der Einfluss der Lesekenntnisse auf die Mathematikkenntnisse mit dem IALS 1994 für die 2.062 deutschen Befragten geschätzt. Der Regressionskoeffizient b der Lesefähigkeit in der Stichprobe beträgt 0,84. Wir möchten wissen, wie stark die Lesekenntnisse die Mathematikkenntnisse in der Grundgesamtheit beeinflussen. Gesucht ist der unbekannte Regressionskoeffizient der Grundgesamtheit β. Das Statistik-Programm Stata gibt für das 95%ige Konfidenzintervall die Grenzen [0,82;0,86] an. Mit 95%iger Wahrscheinlichkeit überdeckt das Intervall den Regressionskoeffizienten der Grundgesamtheit.

Aufgaben zu Konfidenzintervallen

1. Was sind Konfidenzintervalle und wozu benötigt man diese?
2. Wie verändern sich Konfidenzintervalle bei:
   - Vergrößerung der Standardabweichung in der Grundgesamtheit?
   - Vergrößerung der Vertrauenswahrscheinlichkeit?
   - Vergrößerung des Stichprobenumfangs?
3. Im ALLBUS 1994 wurde bei 1.474 westdeutschen Befragten ein durchschnittliches Monatseinkommen $\bar{x}$ von 1.838,39 DM ermittelt. Die Standardabweichung in der Stichprobe $\hat{\sigma}$ beträgt 1.477,68 DM.
   In welchem Bereich liegt das durchschnittliche Monatseinkommen aller westdeutschen Befragten μ mit 95%iger Wahrscheinlichkeit, in welchem Bereich mit 99%iger Wahrscheinlichkeit?
4. Bei der letzten Umfrage der Forschungsgruppe Wahlen vor der Bundestagswahl 1994 gaben 42,5 % der 1.250 Befragten eine CDU/CSU-Wahlabsicht an. Bitte berechnen Sie das 99%ige Konfidenzintervall für den Anteil der CDU/CSU unter allen Wählern, und interpretieren Sie das Ergebnis!
   In der gleichen Umfrage erhielt die PDS 3,5 % bei den Befragten, die eine Wahlabsicht äußerten. Berechnen Sie auch für diese das 99%ige Konfidenzintervall.
5. Bitte berechnen Sie, wie groß die Stichprobe der Forschungsgruppe Wahlen vor der Bundestagswahl 1994 hätte sein müssen, um der FDP mit einer Genauigkeit von 1 % ihren Wähleranteil unter allen Wählern mit 95%iger Sicherheit schätzen zu können! Verwenden Sie die Angaben aus Tabelle 9.1.

12 Hypothesenprüfung

Konfidenzintervalle zählen zu den *Schätzverfahren*, da mit einer Stichprobe Parameter der Grundgesamtheit geschätzt werden. Ebenso wichtig ist die Frage, wie man Hypothesen über eine unbekannte Grundgesamtheit anhand einer einzigen Stichprobe testen kann.

Die Reihe der Testverfahren ist ebenso vielfältig wie die der Schätzverfahren. An dieser Stelle beschränken wir uns auf Tests für Mittelwerte, Mittelwertdifferenzen und den χ^2-Unabhängigkeitstest.

12.1 Grundlagen

Ausgangspunkt einer Untersuchung ist eine Hypothese, d. h. eine noch nicht bewährte Aussage über einen Ausschnitt der sozialen Realität. So könnten wir die Hypothesen aufstellen, dass Männer besser mit Zahlen umgehen können als Frauen oder die Bildungsbeteiligung von Kindern vom Bildungsniveau der Eltern abhängt.

In der Testtheorie bezeichnet man die Hypothese, die überprüft werden soll, als **Alternativhypothese**. Diese beinhaltet die eigentlich interessierende Aussage. Möchten wir geschlechtsspezifische Unterschiede in den Mathematikkenntnissen prüfen, dann stellt die Aussage „Männer können besser mit Zahlen umgehen als Frauen“ die Alternativhypothese dar. Die allgemeine Bezeichnung für die Alternativhypothese ist H_A, manchmal auch H_1.

Als Gegenstück zur Alternativhypothese wird eine **Nullhypothese** formuliert. Sie verneint den in der Alternativhypothese behaupteten Sachverhalt. Die Nullhypothese zur gerade formulierten Alternativhypothese würde lauten: „Männer können nicht besser mit Zahlen umgehen als Frauen“. Die Nullhypothese behauptet, dass es den in der H_A ausgedrückten Unterschied in den numerischen Fähigkeiten von Männern und Frauen nicht gibt. Für die Nullhypothese wird das Kürzel H_0 verwendet.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2_12,

Der erste Schritt der Hypothesenprüfung besteht in der Formulierung der Alternativhypothese und der dazu konkurrierenden Nullhypothese. Sowohl die H_A als auch die H_0 stellen **Behauptungen über die Grundgesamtheit** dar, also beispielsweise über die Wohnbevölkerung der Bundesrepublik.

Anhand einer Stichprobe soll nun überprüft werden, welche der beiden Aussagen über die Grundgesamtheit, H_A oder H_0, zutrifft. Aufgrund der durch Zufallsschwankungen bedingten Abweichung einer Stichprobe von der Grundgesamtheit kann es jedoch zu zwei *Fehlentscheidungen* kommen.

1. So könnten wir aufgrund der Stichprobe zu der Entscheidung gelangen, dass Männer besser mit Zahlen umgehen können als Frauen, obwohl in der Grundgesamtheit kein Unterschied in den numerischen Fähigkeiten besteht. Die Stichprobendaten weisen also auf das Vorliegen der Alternativhypothese hin, während tatsächlich in der Grundgesamtheit die Nullhypothese gilt. Diesen Fehler bei Übertragung der Stichprobenergebnisse auf die Grundgesamtheit nennt man **Fehler 1. Art** oder **α-Fehler**. Er wird als Prozentwert oder als Wahrscheinlichkeit ausgedrückt und auch als *Irrtumswahrscheinlichkeit* bezeichnet.
2. Zum anderen könnten wir auf Basis der Stichprobe zu der Entscheidung gelangen, dass geschlechtsspezifische Unterschiede im Umgang mit Zahlen nicht existieren, obwohl Männer tatsächlich – in der Grundgesamtheit – bessere numerische Fähigkeiten haben als Frauen. Entscheiden wir uns aufgrund der Stichprobe für die Nullhypothese, obwohl in der Grundgesamtheit die Alternativhypothese vorliegt, dann begehen wir den **Fehler 2. Art** bzw. den **β-Fehler**.

Die richtige Entscheidung treffen wir, wenn die aus der Stichprobe gefolgerte Entscheidung mit der Grundgesamtheit übereinstimmt. In Tabelle 12.1 sind alle möglichen Entscheidungen aufgeführt.

Da wir die Grundgesamtheit nicht kennen, können wir nicht mit Sicherheit sagen, ob die Entscheidung auf Basis der Stichprobe richtig ist oder falsch. Wir können aber die Wahrscheinlichkeit des α- und β-Fehlers berechnen.

α-Fehler

Um den α-Fehler berechnen zu können, müssen wir die Verteilung der Stichprobenkennwerte angeben, wenn in der Grundgesamtheit die Nullhy-

Tabelle 12.1: Fehler bei der Hypothesenprüfung

		In der *Grundgesamtheit* gilt:	
		H_0	H_A
Entscheidung aufgrund der **Stichprobe**:	H_0	richtig	β-Fehler
	H_A	α-Fehler	richtig

pothese gilt. Durch den Zentralen Grenzwertsatz wissen wir beispielsweise, dass sich Mittelwerte in Stichproben normal verteilen. Der Mittelwert der Verteilung ist bei Gültigkeit der H_0 der durch die H_0 postulierte Parameter. Die Breite der Mittelwerteverteilung lässt sich aus der Stichprobe schätzen. Ist die Wahrscheinlichkeit für einen vorliegenden Stichprobenkennwert oder einen noch weiter von der H_0 abweichenden Kennwert gering, so wird die Nullhypothese verworfen.

In der Wissenschaft haben sich Grenzen eingebürgert, ab wann ein Stichprobenkennwert bei Gültigkeit der H_0 als unwahrscheinlich einzustufen ist. Die Grenzen befinden sich bei $\alpha = 5\,\%$ bzw. $\alpha = 1\,\%$. Ist die Wahrscheinlichkeit für den ermittelten Stichprobenkennwert bei Gültigkeit der H_0 geringer als 5 % bzw. 1 %, wird die H_0 abgelehnt.

Bei dieser Entscheidung können wir uns irren. Die Wahrscheinlichkeit für den ermittelten Stichprobenkennwert bei Gültigkeit der H_0 ist zwar geringer als 5 % bzw. 1 % – wir können jedoch nicht ausschließen, dass wir eine sehr weit von der Grundgesamtheit abweichende Stichprobe gezogen haben. α gibt daher die Wahrscheinlichkeit an, mit der wir uns bei Ablehnung der Nullhypothese irren (*Irrtumswahrscheinlichkeit*).

β-Fehler

Bei der *Ermittlung des β-Fehlers* lautet die Frage: Wie wahrscheinlich ist das Stichprobenergebnis, wenn in der Grundgesamtheit die Alternativhypothese gilt? Hier ermitteln wir die Verteilung der Stichprobenkennwerte bei Gültigkeit der Alternativhypothese. Ist die Wahrscheinlichkeit des Stichprobenkennwerts bei Gültigkeit der H_A gering, so verwerfen wir die

Alternativhypothese. Auch bei der Ablehnung der Alternativhypothese können wir uns irren, d. h. die Alternativhypothese zu Unrecht ablehnen. Die Wahrscheinlichkeit für unser Stichprobenergebnis oder eine noch größere Abweichung von der H_A gibt die Größe des β-Fehlers an. Ein β-Fehler wird als akzeptabel angesehen, wenn er kleiner als 20 % ist.

α-Fehler und β-Fehler verhalten sich gegenläufig. Je kleiner der α-Fehler, umso größer der β-Fehler. Je sicherer wir sein wollen, dass wir die Nullhypothese nicht zu Unrecht ablehnen, umso eher lehnen wir die Alternativhypothese zu Unrecht ab.

In der Regel wird allerdings nur versucht, die Wahrscheinlichkeit des $\boldsymbol{\alpha}$-Fehlers gering zu halten. Denn um die Kennwerteverteilung angeben zu können, muss man eine präzise Annahme über die Parameter der Grundgesamtheit machen. Dies ist bei der Nullhypothese sehr einfach: sie besagt nämlich normalerweise, dass der Zusammenhang bzw. der Unterschied null ist. In der Alternativhypothese kommt dagegen meist nur eine unpräzise Annahme zum Ausdruck. So wird ein Zusammenhang postuliert, nicht aber, wie groß dieser Zusammenhang ist, oder ein Unterschied wird angenommen, aber nicht, wie groß dieser Unterschied ist. Um den β-Fehler zu testen, müssten wir im obigen Beispiel aber exakt angeben, wie groß der Unterschied in den numerischen Fähigkeiten von Frauen und Männern ist. Wir werden uns im Weiteren mit der Ermittlung des α-Fehlers beschäftigen.

12.2 Test eines Mittelwerts

Bei einem Hypothesentest wird in folgenden Schritten vorgegangen:

1. Null- und Alternativhypothese formulieren und Signifikanzniveau festlegen;
2. Prüfgröße (z. B. z-Wert, t-Wert oder χ^2-Wert) und Verteilung der Prüfgröße bestimmen;
3. Ablehnungsbereich der Nullhypothese kennzeichnen;
4. Prüfgröße berechnen und die Entscheidung über die Nullhypothese treffen.

Zunächst werden diese Schritte für den Test eines Mittelwerts (bei bekannter Varianz in der Grundgesamtheit σ) dargestellt.

1. Null- und Alternativhypothese formulieren, Signifikanzniveau festlegen

Wir möchten wissen, ob die Studienzeit in der Politikwissenschaft durch eine andere Form der Betreuung während des Studiums verändert wird, d. h. ob sie sich verlängert oder verkürzt. Wie diese andere Form der Betreuung aussieht, interessiert uns jetzt nicht näher. Um dies zu prüfen, wurden für ein Pilotprojekt 35 Studierende ausgewählt, die anders betreut wurden als die übrigen Studierenden. Unsere Alternativhypothese besagt, dass anders betreute Studierende kürzer oder auch länger studieren. Skeptiker behaupten dagegen, dass die Art der Betreuung nichts an der Studiendauer ändert. Die Studiendauer bei bisheriger und neuer Betreuung sei identisch. Diese Behauptung beinhaltet die Nullhypothese.

In diesem Beispiel liegt eine *ungerichtete Alternativhypothese* bzw. eine **zweiseitige Fragestellung** vor, da keine Aussage über die Richtung des Unterschieds getroffen wird. Von einer *gerichteten Alternativhypothese* bzw. einer **einseitigen Fragestellung** würde man dagegen sprechen, wenn etwas über die Richtung des Unterschieds ausgesagt würde. Dies wäre z. B. der Fall, wenn wir behaupten würden, dass im Pilotprojekt betreute Studierende schneller ihr Studium abschließen. Die Frage, ob es sich um eine gerichtete oder eine ungerichtete Hypothese bzw. eine ein- oder zweiseitige Fragestellung handelt, wird später für die Bestimmung des Ablehnungsbereiches wichtig.

Wie wir vom Studentensekretariat erfahren haben, beträgt die durchschnittliche Studiendauer im Fach Politikwissenschaft bisher im Schnitt 13,5 Semester (μ_0) und die Standardabweichung 3,2 Semester (σ). Die Angaben des Studentensekretariates sind die Parameter der Grundgesamtheit bei bisheriger Betreuung. Nicht bekannt ist der Mittelwert der Grundgesamtheit bei anderer Betreuung μ. Mit der H_0 und H_A werden nun unterschiedliche Behauptungen über den unbekannten Parameter μ aufgestellt.

Nach der Nullhypothese dauert das Studium anders betreuter Studierender durchschnittlich genauso lange wie bisher:

$$H_0: \quad \mu = \mu_0 = 13{,}5 \quad \text{Semester.}$$

Die Alternativhypothese, dass anders betreute Studierende nicht so viel oder mehr Zeit als bisher zum Erwerb des Examens benötigen, kann man ausdrücken als:

$$H_A: \quad \mu \quad \neq \quad \mu_0 \quad \neq \quad 13{,}5 \quad \text{Semester.}$$

Getestet wird die H_0. Geprüft wird, ob – bei einer bestimmten Irrtumswahrscheinlichkeit – das ermittelte Stichprobenergebnis (hier: durchschnittliche Studiendauer $\bar{x}$ bei neuer Betreuung im Pilotprojekt) mit der Nullhypothese („wahre" durchschnittliche Studiendauer bei anderer Betreuung beträgt $\mu = 13,5$ Semester) vereinbart werden kann. Spricht das ermittelte Stichprobenergebnis gegen die Nullhypothese, dann verwerfen wir diese zugunsten der Alternativhypothese. Lässt sich das Stichprobenergebnis mit der Nullhypothese vereinbaren, dann lehnen wir die Nullhypothese nicht ab.

Ab welchem Stichprobenergebnis die H_0 verworfen wird, hängt davon ab, welche Irrtumswahrscheinlichkeit (α-Fehler) bei der Entscheidung in Kauf genommen wird. Denn auch bei Gültigkeit der H_0 kann – aufgrund der zufälligen Abweichung der Stichprobe von der Grundgesamtheit – ein Stichprobenergebnis vorkommen, das weit vom Parameter der Grundgesamtheit abweicht (auch wenn dies unwahrscheinlich ist).

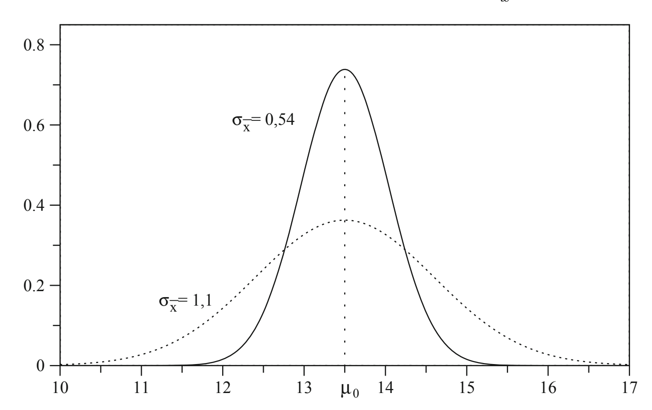
In den Sozialwissenschaften ist es üblich, den α-Fehler nicht größer als 0,05 bzw. 5 % werden zu lassen. Man kann sich dann zu (mindestens) $1 - \alpha = 0,95$ bzw. 95 % sicher sein, die Nullhypothese nicht fälschlicherweise zu verwerfen. $1 - \alpha$ gibt die *Vertrauenswahrscheinlichkeit* an. Wenn man ganz sicher gehen will, legt man die Irrtumswahrscheinlichkeit mit maximal 1 % fest und kann sich damit zu (mindestens) 99 % sicher sein, keinen Fehler zu begehen. Diese Grenzwerte für die Irrtumswahrscheinlichkeit (α-Fehler) werden auch als *Signifikanzniveau* bezeichnet. Die Irrtums- bzw. Vertrauenswahrscheinlichkeiten von 5 % und 1 % bzw. 95 % und 99 % werden in der Wissenschaft als ausreichend betrachtet. Ist die Irrtumswahrscheinlichkeit kleiner als 5 %, spricht man von einem *signifikanten Ergebnis*, ist sie kleiner als 1 %, spricht man von einem *sehr signifikanten* Ergebnis. Üblicherweise kennzeichnet man bei der Darstellung von Ergebnissen die statistischen Kennwerte mit *, wenn sie signifikant sind, und mit **, wenn sie sehr signifikant sind.

Unsere Hypothese soll auf einem Signifikanzniveau von 5 % getestet werden. Welches Signifikanzniveau gewählt wird, hängt von den Konsequenzen ab, die mit einer falschen Entscheidung verbunden sind.

2. Prüfgröße und Verteilung der Prüfgröße bestimmen

Nach dem *Zentralen Grenzwertsatz* verteilen sich Stichprobenmittelwerte normal um den Mittelwert der Grundgesamtheit μ mit einem Standardfehler von $\sigma_{\bar{x}}$ bei hinreichend großen Stichproben (vgl. Kapitel 10.3). Bei Gültigkeit der H_0 ($\mu = \mu_0 = 13,5$ Semester) verteilen sich die Stichprobenmittelwerte normal um $\mu_0 = 13{,}5$ Semester mit dem Standardfehler $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 3,2/\sqrt{35} = 0,54$. Diese Verteilung ist in Abbildung 12.1 (durchgezogene Linie) dargestellt.

Abbildung 12.1: Stichprobenmittelwerteverteilungen mit $\mu_0 = 13,5$ und unterschiedlichen Standardfehlern $\sigma_{\bar{x}}$



Wie man Abbildung 12.1 (durchgezogene Linie) entnehmen kann, ist eine durchschnittliche Studiendauer von z. B. $\bar{x} = 15$ Semestern sehr unwahrscheinlich, *wenn* die durchschnittliche Studiendauer in der Grundgesamtheit 13,5 Semester und der Standardfehler 0,54 Semester beträgt. Erzielen wir in der Stichprobe einen Mittelwert $\bar{x}$, der weit von μ_0 abweicht, so deutet dies darauf hin, dass der Stichprobenmittelwert nicht aus einer Grundgesamtheit stammt, in der die Nullhypothese gilt.

Ob eine bestimmte Abweichung $\bar{x} - \mu_0$ wahrscheinlich ist oder nicht, hängt vom Standardfehler des Mittelwerts $\sigma_{\bar{x}}$ ab. Wie wir wissen, wird $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ mit zunehmendem Stichprobenumfang n und kleiner werdender Standardabweichung des Merkmals in der Grundgesamtheit σ kleiner, d. h. die Verteilung der Stichprobenmittelwerte wird dann schmaler. Die Stichprobenwerte liegen dann näher am Parameter der Grundgesamtheit. In Abbildung 12.1 ist neben der Verteilung, die im Beispiel vorliegt ($\mu_0 = 13,5$, $\sigma_{\bar{x}} = 0,54$), eine weitere Verteilung (gestrichelte Linie) eingezeichnet, die einen größeren Standardfehler ($\sigma_{\bar{x}} = 1,1$) aufweist. Eine durchschnittliche Studiendauer von $\bar{x} = 15$ Semestern (bei $\mu_0 = 13,5$ Semester) ist bei einem Standardfehler des Mittelwerts von $1,1$ viel wahrscheinlicher als bei einem Standardfehler von $0,54$.

Bei der Berechnung der *Prüfgröße* wird deshalb die Abweichung des Stichprobenmittelwerts vom Mittelwert der Grundgesamtheit $\bar{x} - \mu_0$ am Standardfehler des Mittelwerts $\sigma_{\bar{x}}$ relativiert. Die Prüfgröße ist bei diesem Test also einfach der z-transformierte Stichprobenmittelwert $\bar{x}$:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} . \qquad (12.1)$$

Die Verteilung der Prüfgröße wird auch als Testverteilung bezeichnet, da an ihr die H_0 geprüft wird.

3. Ablehnungsbereich der Nullhypothese kennzeichnen

Die Nullhypothese wird abgelehnt, wenn die Wahrscheinlichkeit des Stichprobenkennwerts (Prüfgröße) oder eines noch extremeren Ergebnisses bei Gültigkeit der Nullhypothese gering ist. Gering heißt: Die Wahrscheinlichkeit darf nicht größer als das vorgegebene Signifikanzniveau α, also in der Regel 1 % oder 5 %, werden. Je unwahrscheinlicher der Wert einer Prüfgröße bei Gültigkeit der H_0, umso weiter liegt er in der Testverteilung am Rand.

Bei einer *zweiseitigen* Fragestellung entspricht die Irrtumswahrscheinlichkeit α der Fläche, die an den beiden Rändern der Verteilung der Prüfgröße (hier: der z-Verteilung) liegt (vgl. auch Abbildung 11.2, S. 258). Je nachdem, ob es sich um eine ein- oder zweiseitige Fragestellung handelt, wird

die Fläche entweder nur am linken/rechten Rand oder an beiden Rändern der Testverteilung betrachtet.

Gesucht sind nun die zu dieser Fläche (zu einem bestimmten Signifikanzniveau) gehörenden Grenzwerte der Testverteilung, die als *kritische Werte* (zweiseitige Fragestellung) bzw. als *kritischer Wert* (einseitige Fragestellung) bezeichnet werden. Sie grenzen den Ablehnungsbereich der Nullhypothese vom Nicht-Ablehnungsbereich ab. Da die Prüfgröße z standardnormalverteilt ist, entnehmen wir die *kritischen Werte der Standardnormalverteilung*. Bei der Standardnormalverteilung bzw. z-Verteilung schneidet z. B. der Wert $-1{,}65$ 5 % der Fläche *am linken Rand* ab. Die kritischen Werte der Standardnormalverteilung sind in der Box auf der nächsten Seite dargestellt.

Bei Prüfgrößen, die anders verteilt sind – z. B. χ^2- oder t-verteilt –, müssen die entsprechenden Verteilungen herangezogen werden. Die kritischen Werte für die χ^2- oder t-Verteilung können anhand der Tabellen in Anhang A ermittelt werden.

Die **kritischen Werte der Standardnormalverteilung** lauten:

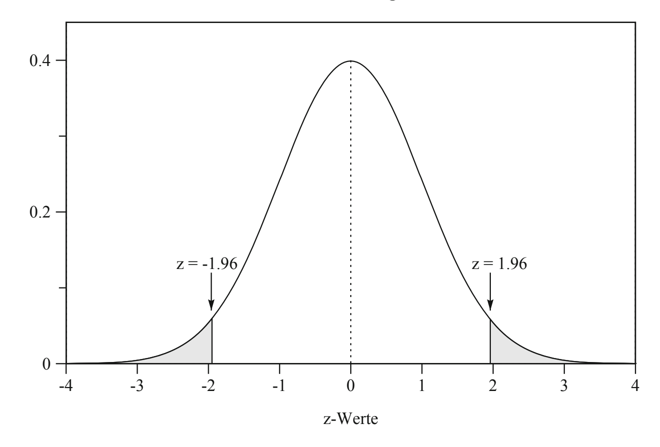
- einseitige Fragestellung
  - 5% Irrtumswahrscheinlichkeit links — $-1{,}65$
    Ablehnungsbereich also: $-\infty$ bis $-1{,}65$
  - 1% Irrtumswahrscheinlichkeit links — $-2{,}33$
    Ablehnungsbereich also: $-\infty$ bis $-2{,}33$
  - 5% Irrtumswahrscheinlichkeit rechts — $1{,}65$
    Ablehnungsbereich also: $1{,}65$ bis ∞
  - 1% Irrtumswahrscheinlichkeit rechts — $2{,}33$
    Ablehnungsbereich also: $2{,}33$ bis ∞
- zweiseitige Fragestellung
  - 5% Irrtumswahrscheinlichkeit — $-1{,}96$ und $1{,}96$
    Ablehnungsbereich also: $-\infty$ bis $-1{,}96$ und $1{,}96$ bis ∞
  - 1% Irrtumswahrscheinlichkeit — $-2{,}58$ und $2{,}58$
    Ablehnungsbereich also: $-\infty$ bis $-2{,}58$ und $2{,}58$ bis ∞

Im gewählten Beispiel liegt eine ungerichtete Alternativhypothese ($\mu \neq \mu_0$) und damit eine *zweiseitige Fragestellung* vor. Die Irrtumswahrscheinlichkeit hatten wir mit 5 % angesetzt. Der Standardnormalverteilung entnehmen wir die Grenzwerte $z_{0,025} = -1{,}96$ und $z_{0,975} = 1{,}96$, d. h. nur in 5 % der Stichproben erhalten wir einen z-Wert, der kleiner als $-1{,}96$ oder größer als 1,96 ist. Ist die auf Basis der Stichprobe der betreuten Studierenden berechnete Prüfgröße kleiner als $-1{,}96$ oder größer als 1,96, lehnen wir die Nullhypothese ab. Nimmt die Prüfgroße dagegen einen Wert zwischen $-1{,}96$ und 1,96 an, dann lehnen wir die Nullhypothese nicht ab.

$$z > |1,96| \longrightarrow H_0 \text{ ablehnen}$$
$$z \leq |1,96| \longrightarrow H_0 \textbf{ nicht} \text{ ablehnen}$$

In Abbildung 12.2 ist der Ablehnungsbereich durch die grau schraffierte Fläche dargestellt.

Abbildung 12.2: Zweiseitiger Ablehnungsbereich (grau schraffierte Fläche) bei einem Signifikanzniveau von 5 % in der Standardnormalverteilung



4. Prüfgröße berechnen und Entscheidung über die Nullhypothese treffen

Die 35 im Rahmen der Pilotstudie anders betreuten Studierenden studierten im Durchschnitt 12 Semester, also 1,5 Semester weniger als bei bisheriger Betreuung. Wir berechnen nun die Prüfgröße, um angeben zu können, an welcher Stelle der Standardnormalverteilung der in der Stichprobe ermittelte $\bar{x}$-Wert liegt:

$$z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{12 - 13{,}5}{\frac{3{,}2}{\sqrt{35}}} = -2{,}77\,.$$

Fällt die Prüfgröße in den Ablehnungsbereich der Nullhypothese, kann diese abgelehnt bzw. „verworfen“ werden. Bei Gültigkeit der Nullhypothese ist es dann sehr unwahrscheinlich, eine Stichprobe mit der beobachteten Prüfgröße zu erhalten. Fällt die Prüfgröße dagegen in den „Annahmebereich“, kann die Nullhypothese nicht verworfen werden.

Da $-2{,}77$ kleiner als $-1{,}96$ ist, wird die Nullhypothese verworfen. Der in der Stichprobe ermittelte Wert ist damit *signifikant*.

Hätten wir in der Stichprobe einen z-Wert zwischen $-1,96$ und $1,96$ ermittelt, dann würden wir die Nullhypothese *nicht ablehnen*. Das heißt jedoch *nicht*, dass wir dann die Nullhypothese *annehmen* könnten. Außer der H_0 kommen auch andere Parameter der Grundgesamtheit als Erzeuger des Stichprobenmittelwerts in Frage. Genau diese Information liefern uns Konfidenzintervalle.

Konfidenzintervall und Signifikanztest

Das Konfidenzintervall für einen Stichprobenmittelwert $\bar{x}$ berechnet sich bei bekannter Standardabweichung in der Grundgesamtheit σ nach Gleichung 11.5 (vgl. Kapitel 11, S. 260). In der Stichprobe der anders betreuten Studierenden wurde eine durchschnittliche Studiendauer von 12 Semestern ($\bar{x}$) festgestellt. Gefragt wird, wo die durchschnittliche Studiendauer μ bei anderer Betreuung in der Grundgesamtheit liegt. Bei einer Vertrauenswahrscheinlichkeit von 95 % (bzw. einer Irrtumswahrscheinlichkeit von 5 %) resultiert

$$\begin{aligned} 12 - 1{,}96 \cdot \frac{3,2}{\sqrt{35}} \quad &\leq \mu \leq \quad 12 - 1{,}96 \cdot \frac{3,2}{\sqrt{35}} \\ 10{,}94 \quad &\leq \mu \leq \quad 13{,}06\,. \end{aligned}$$

Das Intervall [10,94; 13,06] überdeckt mit 95%iger Wahrscheinlichkeit die durchschnittliche Studiendauer in der Grundgesamtheit. Das Konfidenzintervall umschließt also nicht den im Signifikanztest durch die H_0 postulierten Wert von 13, 5 Semestern.

Wird die H_0 in einem Signifikanztest bei einem bestimmten Signifikanzniveau α abgelehnt, dann überdeckt das für eine Vertrauenswahrscheinlichkeit von $1 - \alpha$ berechnete Konfidenzintervall auch nicht den von der H_0 postulierten Wert der Grundgesamtheit. Wissen wir auf Basis des Signifikanztests nur, dass (bei gegebener Irrtumswahrscheinlichkeit) der Stichprobenmittelwert (bei neuer Betreuung) von 12 Semestern nicht mit der Nullhypothese vereinbar ist, so gibt uns das Konfidenzintervall zusätzlich die Information, in welchem Bereich die durchschnittliche Studiendauer bei anderer Betreuung in der Grundgesamtheit (bei einer gegebenen Vertrauenswahrscheinlichkeit) liegt.

Einseitige Fragestellung

Eine einseitige Fragestellung liegt vor, wenn die Alternativhypothese lautet, dass die neue Betreuungsform die Studienzeit verkürzt. Dem Inhalt der Nullhypothese entspricht dann die These, dass die Studiendauer bei alternativer Betreuung gleich bleibt oder zunimmt.

$$H_0: \quad \mu \geq \mu_0 \geq 13{,}5 \quad \text{und} \quad H_A: \quad \mu < \mu_0 < 13{,}5$$

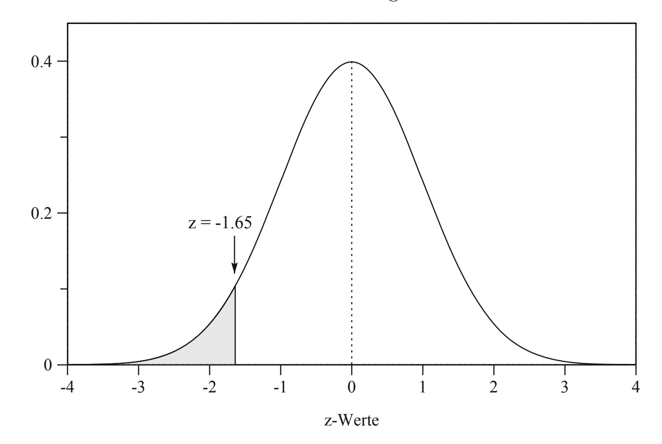
Da die Alternativhypothese eine Verkürzung der Studiendauer postuliert, muss der gesamte Ablehnungsbereich links liegen. Bei einer Irrtumswahrscheinlichkeit von 5 % ergibt sich ein kritischer Wert von $z = -1,65$ (vgl. S. 280). Ist der in der Stichprobe ermittelte z-Wert kleiner als $-1,65$, dann wird die H_0 abgelehnt. Ist er größer als $-1,65$, dann wird die H_0 nicht abgelehnt (vgl. zur Kennzeichnung des Ablehnungsbereiches bei einseitiger Fragestellung auch die Ausführungen auf Seite 295).

$z < -1,65 \quad \longrightarrow \quad H_0$ ablehnen

$z \geq -1,65 \quad \longrightarrow \quad H_0$ **nicht** ablehnen

Der Ablehnungsbereich ist in Abbildung 12.3 grau schraffiert dargestellt.

Abbildung 12.3: Einseitiger Ablehnungsbereich (grau schraffierte Fläche) bei einem Signifikanzniveau von 5 % in der Standardnormalverteilung



Die Prüfgröße beträgt $z = -2,77$, wie wir auf S. 282 bereits berechnet haben. Da die Prüfgröße $z = -2,77$ kleiner als der kritische Wert $z = -1,65$ ist, wird die Nullhypothese verworfen. Die Studiendauer mit neuer Betreuung unterscheidet sich also signifikant von der Studiendauer bei alter Betreuung.

Wie man an den Ablehnungsbereichen in Abbildung 12.2 und Abbildung 12.3 sieht, liegen bei einer einseitigen Fragestellung (gerichtete Alternativhypothese) schon geringere Abweichungen in Richtung der Alternativhypothese im Ablehnungsbereich als bei einer zweiseitigen Fragestellung (ungerichtete Alternativhypothese). Ob eine gerichtete oder ungerichtete Alternativhypothese formuliert wird, muss vor der Berechnung der Prüfgröße festgelegt werden. Es kann nämlich passieren, dass eine Prüfgröße bei

einseitiger Fragestellung, nicht aber bei zweiseitiger Fragestellung signifikant ist. Im Nachhinein eine gerichtete Alternativhypothese aufzustellen (damit eine Prüfgröße in den Ablehnungsbereich der Nullhypothese fällt und signifikant wird), ist nicht seriös.

p-Werte - empirisches Signifikanzniveau

Man kann die Irrtumswahrscheinlichkeit für den in der Stichprobe ermittelten Wert der Prüfgröße auch genau bestimmen. Für die *einseitige* Fragestellung im Beispiel entnimmt man der z-Tabelle die Fläche, die links vom Wert der Prüfgröße $-2{,}77$ liegt, nämlich $\approx 0{,}0028$. Die Wahrscheinlichkeit, in der Stichprobe einen z-Wert kleiner als $-2{,}77$ zu erhalten, wenn in der Grundgesamtheit die H_0 gilt, beträgt also $0,28\,\%$. Die Irrtumswahrscheinlichkeit für die *zweiseitige Fragestellung* beträgt im Beispiel $\approx 0{,}0056$, da hier der Ablehnungsbereich der Nullhypothese auf beiden Seiten der Verteilung liegt. Es muss also die Fläche, die links von $-2,77$ liegt, zur Fläche, die sich rechts von $+2,77$ befindet, addiert werden $(0,0028 + 0,0028 = 0,0056)$. Die Wahrscheinlichkeit in der Stichprobe einen z-Wert zu erhalten, der kleiner als $-2,77$ oder größer als $+2,77$ ist, wenn in der Grundgesamtheit die H_0 gilt, beträgt $0,56\,\%$.

Diese für die Prüfgröße berechnete „empirische" Irrtumswahrscheinlichkeit wird auch als *p-Wert* bezeichnet. Der p-Wert gibt die *Wahrscheinlichkeit an, bei Gültigkeit der H_0 den Wert der Prüfgröße oder einen mit der H_0 noch weniger zu vereinbarenden Wert in der Stichprobe zu erhalten.* Die meisten Statistikprogramme geben p-Werte an. Ist der p-Wert kleiner als das gewählte Signifikanzniveau α, dann wird die Nullhypothese verworfen, ist der p-Wert größer, dann wird die Nullhypothese nicht verworfen.

$$
\begin{aligned}
p\text{-Wert} &< \; 0,05 \,(\text{bzw. } 0,01) && \longrightarrow && H_0 \,\text{ablehnen} \\
p\text{-Wert} &\geq \; 0,05 \,(\text{bzw. } 0,01) && \longrightarrow && H_0 \,\textbf{nicht}\text{ ablehnen}
\end{aligned}
$$

Statistik-Programme geben häufig zweiseitige p-Werte an. Hier wäre dies $p = 0,0056$. Liegt eine einseitige Fragestellung vor, muss der zweiseitige p-Wert halbiert werden $(p = 0,0056/2 = 0,0028)$.

Test eines Mittelwerts bei unbekanntem σ

Ist die Standardabweichung der Grundgesamtheit σ nicht bekannt, dann wird diese durch die Standardabweichung der Stichprobe $\hat{\sigma}$ geschätzt (vgl. Gleichung 11.6, S. 263), damit der Standardfehler des arithmetischen Mittels $\hat{\sigma}_{\bar{x}}$ bestimmt werden kann. Die Prüfgröße (vgl. auch Gleichung 12.1) ist in diesem Fall (mit $df = n-1$ Freiheitsgraden) t- und nicht z-verteilt.

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}} \tag{12.2}$$

Da die t-Verteilung breiter als die z-Verteilung ist, sind die kritischen Werte hier größer. Bei großen Stichproben nähern sich die kritischen Werte beider Verteilungen an. Bei Stichproben kleiner als $n = 30$ müssen die Merkmale in der Grundgesamtheit normalverteilt sein.

12.3 Tests für Mittelwertunterschiede

Mit diesen Tests werden Hypothesen über Mittelwertunterschiede in der Grundgesamtheit überprüft. Eine These wäre, dass der Umgang mit Zahlen Männern leichter fällt als Frauen. Untersuchen können wir die These mit den Daten des IALS 1994, die bereits in Kapitel 7.5 verwendet wurden. Wir beschränken uns auch hier auf den deutschen Teil des IALS. Männer und Frauen können als zwei Stichproben aufgefasst werden.

Eine wichtige Frage für die Auswahl ist, ob es sich um *unabhängige* oder *abhängige* Stichproben handelt. Bei unabhängigen Stichproben beeinflusst die Auswahl der Elemente der einen Stichprobe die Auswahl der Elemente der anderen Stichprobe *nicht*. Bei Männern und Frauen im IALS handelt es sich um unabhängige Stichproben, weil die Auswahl von Männern und Frauen unabhängig voneinander erfolgte. Bei abhängigen Stichproben beeinflusst die Auswahl der Elemente der einen Stichprobe die Auswahl der Elemente der anderen Stichprobe. Ein typisches Beispiel sind Wiederholungsmessungen. Werden Statistikkenntnisse von Studierenden *vor* und *nach dem Besuch* eines Statistikkurses erhoben, so liegen Messungen zu zwei Zeitpunkten vor. Die Auswahl der Studierenden zum Zeitpunkt 1 (Stichprobe 1) bestimmt die Studierenden, die zum Zeitpunkt 2 untersucht werden (Stichprobe 2). Abhängige Stichproben liegen auch dann

vor, wenn in einer Studie Ehepaare untersucht werden. Männer und Frauen einer Stichprobe von Ehepartnern können wiederum als zwei getrennte Stichproben behandelt werden. Männer und Frauen werden hier jedoch nicht unabhängig voneinander ausgewählt. Die beiden Stichproben sind abhängig (vgl. Kapitel 12.3.2).

12.3.1 Test für unabhängige Stichproben

Die Mathematikkenntnisse der im deutschen Teil des IALS befragten 938 Männer belaufen sich auf durchschnittlich $\bar{x}_1 = 296,14$ Punkte. Die 1124 in Deutschland befragten Frauen erzielen durchschnittlich $\bar{x}_2 = 288,14$ Punkte. In der Stichprobe schneiden die Männer demnach etwas besser ab als die Frauen, $\bar{x}_1 - \bar{x}_2 = 296,14 - 288,14 = 8$ Punkte. Die eigentlich interessierende Frage ist, ob sich die Mathematikkenntnisse von Männern und Frauen in der Grundgesamtheit unterscheiden.

1. Null- und Alternativhypothese formulieren, Signifikanzniveau festlegen

Die Alternativhypothese H_A postuliert in diesem Fall, dass ein Unterschied in den numerischen Fähigkeiten von Männern und Frauen bestehen. Mit der H_0 behaupten wir, dass es keinen Unterschied zwischen Männern und Frauen gibt. Wir formulieren also eine ungerichtete Alternativhypothese.

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \quad &\text{bzw.} \quad \mu_1 - \mu_2 = 0 \\ H_A: &\quad \mu_1 \neq \mu_2 \quad &\text{bzw.} \quad \mu_1 - \mu_2 \neq 0 \end{aligned}$$

Die Irrtumswahrscheinlichkeit setzen wir mit 5 % fest. Die Wahrscheinlichkeit, die H_0 abzulehnen, obwohl diese in der Grundgesamtheit gilt, soll maximal 5 % betragen.

2. Prüfgröße und Verteilung der Prüfgröße bestimmen

Weil sich Stichprobenmittelwerte $\bar{x}$ nach dem Zentralen Grenzwertsatz normalverteilen, sind Mittelwertdifferenzen unabhängiger Stichproben $\bar{x}_1 - \bar{x}_2$ in hinreichend großen Stichproben ebenfalls normalverteilt. Der

Erwartungswert der Stichprobenmittelwerteverteilung ist die Mittelwertdifferenz in der Grundgesamtheit $\mu_1 - \mu_2$. Auch hier wird vorausgesetzt, dass die Stichprobenumfänge n_1 und n_2 hinreichend groß sind, also $n_1 > 30$ und $n_2 > 30$. Ist $n_1 \leq 30$ oder $n_2 \leq 30$, dann müssen die Merkmale in der *Grundgesamtheit* normalverteilt sein, damit die Mittelwertdifferenzen in Stichproben normalverteilt sind. Der Standardfehler von $\bar{x}_1 - \bar{x}_2$ wird mit $\sigma_{(\bar{x}_1 - \bar{x}_2)}$ bezeichnet. Er gibt an, wie stark die Mittelwertdifferenzen in Stichproben von der Mittelwertdifferenz der Grundgesamtheit abweichen.

Auch hier ist eine konkrete Abweichung $(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)$ bei einer breiten Kennwerteverteilung (großer Standardfehler) wahrscheinlicher als bei einer schmalen Kennwerteverteilung (kleiner Standardfehler). Die *Prüfgröße z* standardisiert die Abweichung der Mittelwertdifferenz $\bar{x}_1 - \bar{x}_2$ von der durch die H_0 postulierten Mittelwertdifferenz der Grundgesamtheit $\mu_1 - \mu_2$, indem durch den Standardfehler $\sigma_{(\bar{x}_1 - \bar{x}_2)}$ dividiert wird.

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} \tag{12.3}$$

Da bei Gültigkeit der gewählten Nullhypothese $\mu_1 - \mu_2 = 0$ ist, vereinfacht sich die Formel zu:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{(\bar{x}_1 - \bar{x}_2)}} \, . \tag{12.4}$$

Der *Standardfehler der Mittelwertdifferenz* $\sigma_{(\bar{x}_1 - \bar{x}_2)}$ berechnet sich aus den Varianzen des Merkmals für beide Gruppen in der Grundgesamtheit:

$$\sigma_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \, . \tag{12.5}$$

Sind die Varianzen σ_1^2 und σ_2^2 in der Grundgesamtheit unbekannt, werden diese durch die Stichprobe geschätzt: $\hat{\sigma}_1^2 = SAQ/(n_1 - 1)$ und $\hat{\sigma}_2^2 = SAQ/(n_2 - 1)$).

$$\hat{\sigma}_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \tag{12.6}$$

Setzt man Gleichung 12.6 in Gleichung 12.4 ein, erhält man als *Prüfgröße*:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}_{(\bar{x}_1-\bar{x}_2)}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}\,. \tag{12.7}$$

Da die Prüfgröße bei großen Stichproben standardnormalverteilt ist, können die kritischen Werte der z-Tabelle entnommen werden (vgl. dazu auch den Abschnitt zu t-Tests auf S. 291).

3. Ablehnungsbereich der Nullhypothese kennzeichnen

Bei einem Signifikanzniveau von $\alpha = 0,05 = 5\,\%$, einer zweiseitigen Fragestellung und einer z-verteilten Prüfgröße müssen die beiden Werte ermittelt werden, die links und rechts von der z-Verteilung jeweils 0,025 bzw. 2,5 % der Gesamtfläche abschneiden. Aus der z-Tabelle entnimmt man für die untere Grenze $z_{0{,}025}$ den Wert −1,96 und für die obere Grenze $z_{0{,}975}$ den Wert +1,96. Gilt die H_0, dann ist die Wahrscheinlichkeit, einen z-Wert in der Stichprobe kleiner als −1,96 oder größer als 1,96 zu erhalten, kleiner als 5 %. Wir lehnen die Nullhypothese daher ab, wenn die Prüfgröße kleiner als −1,96 oder größer als 1,96 ist. Die Nullhypothese wird nicht verworfen, wenn die Prüfgröße zwischen −1,96 und 1,96 liegt.

4. Prüfgröße berechnen und Entscheidung über die Nullhypothese treffen

Die aus den Stichproben geschätzte Varianz der Mathematikkenntnisse beträgt für Männer $\hat{\sigma}_1^2 = 2245$ und für Frauen $\hat{\sigma}_1^2 = 1858$. Setzt man die Stichprobenwerte in Gleichung 12.7 ein, dann erhält man:

$$z = \frac{296,14 - 288,14}{\sqrt{\frac{2245}{938} + \frac{1858}{1124}}} = 3{,}98\,.$$

Da $3,98$ größer als $+1{,}96$ ist, kann die Nullhypothese verworfen werden. Der Unterschied in den Mathematikkenntnissen von Männern und Frauen ist signifikant. Weil die Prüfgröße auch kleiner als der zu einer Irrtumswahrscheinlichkeit von 1 % gehörende kritische Wert von +2,58 ist, kann man den Unterschied auch als „sehr signifikant" bezeichnen. Auch bei einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}01$ lehnen wir die Nullhypothese ab.

Konfidenzintervall

Auch hier soll zum Vergleich das Konfidenzintervall berechnet werden. Das Konfidenzintervall für Mittelwertdifferenzen (bei hinreichend großen Stichproben) lässt sich ganz einfach bestimmen.

$$\underbrace{(\bar{x}_1 - \bar{x}_2) - z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{(\bar{x}_1-\bar{x}_2)}}_{\text{untere Grenze}} \quad \leq \mu_1 - \mu_2 \leq \quad \underbrace{(\bar{x}_1 - \bar{x}_2) + z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{(\bar{x}_1-\bar{x}_2)}}_{\text{obere Grenze}} \tag{12.8}$$

$\bar{x}_1 - \bar{x}_2$ beträgt + 8, der Standardfehler $\hat{\sigma}_{(\bar{x}_1-\bar{x}_2)}$ wird durch Gleichung 12.6 bestimmt. Setzt man die Werte ein, resultiert bei einem z-Wert von 1,96 für eine Vertrauenswahrscheinlichkeit von 95 %

$$\begin{aligned} 8 - 1{,}96 \cdot \sqrt{\frac{2245}{938} + \frac{1858}{1124}} \quad &\leq \mu_1 - \mu_2 \leq \quad 8 + 1{,}96 \cdot \sqrt{\frac{2245}{938} + \frac{1858}{1124}} \\ 4{,}1 \quad &\leq \mu_1 - \mu_2 \leq \quad 11,9\,. \end{aligned}$$

Das Intervall [4,1; 11,9] überdeckt mit 95 %iger Wahrscheinlichkeit die Differenz in den Mathematikkenntnissen von Männern und Frauen. Das Konfidenzintervall umschließt nicht den von der Nullhypothese postulierten Wert ($\mu_1 - \mu_2 = 0$).

Statistische Signifikanz und praktische Bedeutsamkeit

Durch den Test für Mittelwertdifferenzen wurde ein signifikanter Unterschied in den Mathematikkenntnissen von Männern und Frauen festgestellt. Mit hoher Wahrscheinlichkeit verfügen Männer in der Grundgesamtheit über bessere Fähigkeiten im Bereich Alltagsmathematik.

Statistische Signifikanz sollte jedoch *nicht* mit praktischer Bedeutsamkeit verwechselt werden. In Kapitel 7.5 wurde die Stärke des Zusammenhangs zwischen dem Geschlecht und den Kenntnissen in Alltagsmathematik mit η^2 quantifiziert. Lediglich rund 0,8 % der Unterschiede in den Mathematikkenntnissen lassen sich durch das Geschlecht erklären. Zur Erklärung der Mathematikkenntnisse der Befragten im IALS 1994 ist das Geschlecht praktisch bedeutungslos. Zusammenfassend kann man beide Befunde auch so ausdrücken: Die etwas besseren Mathematikkenntnisse der Männer in der Stichprobe bestehen mit hoher Wahrscheinlichkeit auch in der Grundgesamtheit (Deutschland 1994).

Statistische Signifikanz hängt von zwei Faktoren ab: der Stärke des Effekts und der Größe der Stichprobe. Ist die Stichprobe groß, dann ist der Standardfehler klein und die Kennwerteverteilung schmal. Mit sehr großen Stichproben lassen sich deshalb bereits geringe Unterschiede und schwache Zusammenhänge in der Grundgesamtheit nachweisen. Ein statistisch signifikantes Ergebnis muss also inhaltlich nicht bedeutsam sein. Umgekehrt ist ein starker Zusammenhang oder Unterschied in der Stichprobe wenig aussagekräftig, wenn er nicht auf die Grundgesamtheit übertragen werden kann – also statistisch nicht signifikant ist.

t-Tests

Statistikprogramme führen einen exakten t- und nicht den oben angegebenen z-Test durch. Mit einem t-Test wird die Unsicherheit berücksichtigt, die mit der Schätzung des Standardfehlers durch die Stichprobendaten verbunden ist. Vor allem bei kleinen Stichproben ist dies von Bedeutung. Bei großen Stichproben – wie im obigen Beispiel – macht es keinen Unterschied, ob die Ablehnungsbereiche anhand der t- oder der z-Verteilung ermittelt werden. Mit zunehmender Stichprobengröße n nähert sich die t-Verteilung einer z-Verteilung an.

Bei einer t-Verteilung der Prüfgröße müssen die Freiheitsgrade bestimmt werden. Die Ermittlung der Freiheitsgrade bei einer t-Verteilung der Prüfgröße in Gleichung 12.6 ist aufwändig (vgl. Sachs 2006, 340). (Das ist auch der Grund, warum in Statistik-Lehrbüchern approximativ von einer z-Verteilung ausgegangen wird). Stata ermittelt mit einem t-Test (für $\sigma_1^2 \neq \sigma_2^2$) einen t-Wert von $3,97$ bei $df = 1915{,}79$ Freiheitsgraden. Die untere Grenze des von Stata ausgegebenen 95%igen Konfidenzintervalls liegt

bei 4,1, die obere Grenze bei 11,9 –. Die Werte sind also fast identisch mit den oben berechneten.

Statistikprogramme bieten außerdem einen t-Test für Mittelwertvergleiche an, wenn davon ausgegangen werden kann, dass die (unbekannten) Varianzen σ_1^2 und σ_2^2 in der Grundgesamtheit identisch sind. Ob in der Grundgesamtheit gleiche oder ungleiche Varianzen vorliegen, kann durch einen weiteren Test – den F-Test – geprüft werden.[1] Im Beispiel wird die Annahme gleicher Varianzen der Mathematikkenntnisse für Männer und für Frauen in der Grundgesamtheit abgelehnt. Ist die Annahme gleicher Varianzen gerechtfertigt, dann kann $\hat{\sigma}^2$ für beide Gruppen gemeinsam geschätzt werden (Agresti und Finlay 2008, 197). Die Ermittlung der Freiheitsgrade ist bei Annahme von Varianzhomogenität einfach ($df = n_1 + n_2 - 2$).

12.3.2 Test für abhängige Stichproben

Ein typisches Beispiel für abhängige Stichproben sind Wiederholungsmessungen an ein und denselben Personen, wie sie z. B. vorliegen würden, wenn Studierende zu Beginn und zum Ende eines Statistikkurses eine Klausur schreiben würden. Für diese Studierenden hätte man dann jeweils zwei Klausurnoten, die man wiederum als zwei Stichproben auffassen kann – allerdings als *abhängige* oder *gepaarte Stichproben*. Bei abhängigen Stichproben ist jede der beiden Stichproben gleich groß, da der Wert in einer Stichprobe mit einem Wert aus der anderen Stichprobe verbunden sein muss.

Uns interessiert, ob Statistikkurse Statistikkenntnisse verbessern. Zur Überprüfung der Hypothese haben wir aus allen Statistikkursen – das Beispiel ist erfunden – eine Zufallsstichprobe von 32 Studierenden gezogen. Zu Semesterbeginn als auch zum Abschluss des Semesters haben die 32 zufällig ausgewählten Studierenden eine Klausur geschrieben. Als Indikator für die Kenntnisse in Statistik wird die Punktezahl in jeder Klausur herangezogen. Für jeden Studierenden liegt ein Paar von Messwerten vor, für das die Differenz d_i berechnet werden kann:

---

1 Ein F-Test setzt normalverteilte Merkmale in der Grundgesamtheit (!) voraus. Weil die Mathematikkenntnisse von Männern und Frauen in der Stichprobe näherungsweise normalverteilt sind (vgl. Abbildung 7.2), kann im Beispiel eine Normalverteilung in der Grundgesamtheit unterstellt werden. F-Tests sind gegenüber einer Verletzung der Normalverteilungsannahme nicht robust, weshalb Agresti und Finlay (2008, 200) von der Verwendung des F-Tests abraten.

$$d_i = \text{Messwert in Stichprobe 2} - \text{Messwert in Stichprobe 1}\,.$$

i kennzeichnet dabei den Laufindex für die Paare, der im Beispiel von $i = 1$ bis $i = n = 32$ läuft. d_i gibt hier die Differenz zwischen der Anzahl der Punkte in der zweiten und ersten Klausur für einen Studierenden (ein Paar von Messwerten) an. Diese Differenz wird für alle 32 Studierenden ermittelt.

Man kann nun für diese gemessenen Differenzen d_i das arithmetische Mittel $\bar{x}_d$ berechnen und die Standardabweichung für die Grundgesamtheit $\hat{\sigma}_d$ schätzen.

$$\bar{x}_d = \frac{\sum_{i=1}^{n} d_i}{n} \tag{12.9}$$

$$\hat{\sigma}_d = \frac{\sum_{i=1}^{n} (d_i - \bar{x}_d)^2}{n-1} \tag{12.10}$$

Das arithmetische Mittel beträgt im Beispiel 13 Punkte, die geschätzte Standardabweichung 6 Punkte, also $\bar{x}_d = 13$ und $\hat{\sigma}_d = 6$. Durchschnittlich wurden in der zweiten Klausur also 13 Punkte mehr erzielt als in der ersten Klausur. Geprüft werden soll nun, ob aus der ermittelten durchschnittlichen Differenz in der Stichprobe $\bar{x}_d$ auch auf eine durchschnittliche Differenz in der Grundgesamtheit μ_d geschlossen werden kann.

1. Null- und Alternativhypothese formulieren, Signifikanzniveau festlegen

Die Alternativhypothese lautet: „Die Statistikkenntnisse werden durch die Kursteilnahme verbessert.“ Als Nullhypothese formulieren wir: „Die Statistikkenntnisse werden durch die Kursteilnahme nicht verbessert.“ „Nicht verbessert“ kann sowohl „gleich bleiben“ als auch „verschlechtern“ bedeuten. Die Nullhypothese gibt hier also einen Bereich an. Dies ist immer

der Fall, wenn die Alternativhypothese gerichtet ist, also eine einseitige Fragestellung vorliegt.

$$H_0: \quad \mu_d \leq 0$$
$$H_A: \quad \mu_d > 0$$

Mit μ_d wird hier der *Mittelwert der Differenzen* in der Grundgesamtheit bezeichnet (im Unterschied zur *Differenz der Mittelwerte* $\mu_1 - \mu_2$ beim Test für unabhängige Stichproben).

Wir legen die Irrtumswahrscheinlichkeit (Signifikanzniveau) mit $\alpha = 0{,}01 = 1\,\%$ fest, da wir bei einer Ablehnung der Nullhypothese sehr sicher gehen wollen.

Mit der formulierten Nullhypothese $\mu_d \leq 0$ sind mehrere Verteilungen der Grundgesamtheit vereinbar. So könnte in Wahrheit die durchschnittliche Differenz $\mu_d = 0$ sein, sie könnte aber auch $\mu_d = -1$ sein, $\mu_d = -2$ oder $\mu_d = -3{,}85$ etc. betragen, also völlig beliebige Werte kleiner null annehmen. Um die kritischen Werte bestimmen zu können, benötigen wir aber eine konkrete Annahme über μ_d. Welche konkrete Annahme über die Grundgesamtheit bei Gültigkeit dieser H_0 soll nun gemacht werden?

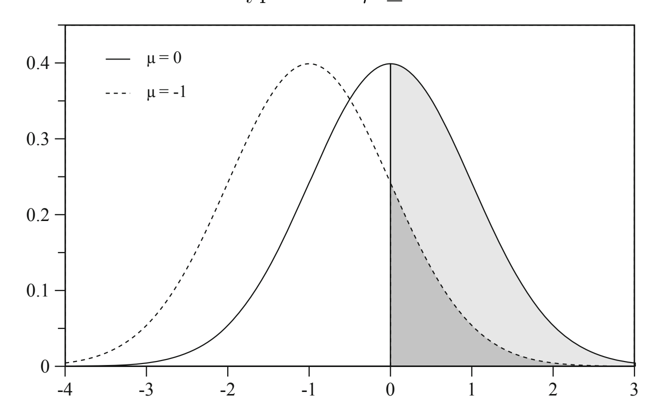
Einseitiger Ablehnungsbereich

Die Lösung des Problems ist einfach und soll ganz allgemein am Beispiel einer Normalverteilung veranschaulicht werden. Wir betrachten die Fläche am rechten Rand der Verteilung, da die Alternativhypothese im Beispiel ja Werte größer als null postuliert und der Ablehnungsbereich der Nullhypothese (in Größe der Irrtumswahrscheinlichkeit α) am rechten Rand der Verteilung liegen muss. Eine Grundgesamtheit, in der $\mu = 0$ gilt, bewirkt in 50 % aller Stichproben einen Mittelwert größer 0. Eine Grundgesamtheit dagegen, in der z. B. $\mu = -1$ gilt, produziert dagegen in 50 % aller Stichproben einen Wert größer als -1. Dies bedeutet, dass die Wahrscheinlichkeit für einen Wert größer 0 bei der Grundgesamtheit mit $\mu = 0$ 50 % beträgt, bei der Grundgesamtheit mit $\mu = -1$ aber geringer als 50 % sein muss.

Abbildung 12.4 zeigt zwei Kennwerteverteilungen, eine basierend auf der Annahme $\mu = 0$ (durchgezogene Linie), die andere für die Annahme

$\mu = -1$ (gestrichelte Linie). Die Wahrscheinlichkeit, bei $\mu = -1$ einen Stichprobenkennwert zu erhalten, der größer als 0 ist, entspricht genau der ganz dunkel schraffierten Fläche. Die ganz dunkle Fläche gibt also die Irrtumswahrscheinlichkeit für Werte größer und gleich null an, wenn in der Grundgesamtheit $\mu = -1$ gilt. Bei $\mu = 0$ entspricht die Irrtumswahrscheinlichkeit für den Wert null aber der *gesamten* schraffierten Fläche. Die Irrtumswahrscheinlichkeit ist für einen gegebenen Stichprobenkennwert in einer Kennwerteverteilung mit $\mu = 0$ größer als für jede Kennwerteverteilung in der $\mu < 0$ ist.

Abbildung 12.4: Irrtumswahrscheinlichkeiten für den Wert 0 bei verschiedenen Nullhypothesen $\mu \leq 0$



Aus diesem Grund wird auch dann, wenn die Nullhypothese einen Bereich angibt, die Nullhypothese $\mu = 0$ getestet.[2] Genau dies haben wir bei dem einseitigen Test eines Mittelwerts bereits getan. Dort wurde tatsächlich die Nullhypothese $\mu = 13,5$ getestet (vgl. S. 283).

2. Prüfgröße und Verteilung der Prüfgröße bestimmen

Bei hinreichend großen Stichproben, d. h. mehr als 30 Messwertpaaren, verteilen sich die arithmetischen Mittel der Differenzen aus Stichproben $\bar{x}_d$

2 Dies gilt natürlich auch für eine Nullhypothese, die einen Wert größer oder gleich null ($\mu \geq 0$) postuliert.

annähernd normal um das arithmetische Mittel der Differenz der Grundgesamtheit μ_d mit einem Standardfehler von $\sigma_{\bar{x}_d}$.

Bei der Berechnung der Prüfgröße wird berücksichtigt, wie groß der Mittelwert der Differenzen $\bar{x}_d$ in der Stichprobe im Vergleich zum Mittelwert der Differenzen in der Grundgesamtheit μ_d (bei Gültigkeit der Nullhypothese) ist. Der Unterschied zwischen dem Mittelwert der Differenzen in der Stichprobe und der Grundgesamtheit $\bar{x}_d - \mu_d$ muss auch hier an der Breite der Kennwerteverteilung, dem *Standardfehler* des Mittelwerts der Differenzen $\sigma_{\bar{x}_d}$, relativiert werden. Ist der Standardfehler $\sigma_{\bar{x}_d}$ groß (z.B. weil der Stichprobenumfang gering ist), dann sind große Abweichungen $\bar{x}_d - \mu_d$ wahrscheinlicher als bei kleinem Standardfehler.

Da der Standardfehler $\hat{\sigma}_{\bar{x}_d}$ auf Basis der Stichprobe geschätzt wird, ist die Prüfgröße (mit $df = n - 1$ Freiheitsgraden) t-verteilt:

$$t = \frac{\bar{x}_d - \mu_d}{\hat{\sigma}_{\bar{x}_d}}\,. \qquad (12.11)$$

Bei der hier gewählten Nullhypothese $\mu_d = 0$, vereinfacht sich die Berechnung zu:

$$t = \frac{\bar{x}_d - 0}{\hat{\sigma}_{\bar{x}_d}} = \frac{\bar{x}_d}{\hat{\sigma}_{\bar{x}_d}}\,. \qquad (12.12)$$

Der Standardfehler des arithmetischen Mittels der Differenzen $\sigma_{\bar{x}_d}$ berechnet sich aus dem Stichprobenumfang n und der Standardabweichung der Differenzen in der Grundgesamtheit.

$$\sigma_{\bar{x}_d} = \frac{\sigma_d}{\sqrt{n}} \qquad (12.13)$$

Wird die Standardabweichung der Grundgesamtheit σ_d wie hier durch die Stichprobe geschätzt $\hat{\sigma}_d$ (wobei zur Berechnung von $\hat{\sigma}_d$ die SAQ_d durch $n - 1$ Messwertpaare dividiert wird), lautet die Gleichung:

$$\hat{\sigma}_{\bar{x}_d} = \frac{\hat{\sigma}_d}{\sqrt{n}}\,. \tag{12.14}$$

Der geschätzte Standardfehler des arithmetischen Mittels der Differenzen $\hat{\sigma}_{\bar{x}_d}$ wird nun in Gleichung 12.12 eingesetzt, womit wir die Prüfgröße berechnen können:

$$t = \frac{\bar{x}_d}{\frac{\hat{\sigma}_d}{\sqrt{n}}}\,. \tag{12.15}$$

3. Ablehnungsbereich der Nullhypothese kennzeichnen

Die Prüfgröße ist mit $n-1$ Freiheitsgraden t-verteilt. Die Irrtumswahrscheinlichkeit wurde mit $\alpha = 0{,}01 = 1\,\%$ festgesetzt. Da die Alternativhypothese eine Verbesserung der Statistikkenntnisse behauptet, muss der gesamte Ablehnungsbereich der Nullhypothese am rechten Rand der t-Verteilung liegen. Wir suchen also den t-Wert, der bei einer Verteilung mit 31 Freiheitsgraden am rechten Ende der Verteilung $0{,}01$ der Fläche abschneidet. In der t-Tabelle im Anhang sind im Kopf der Tabelle die Flächen, die *links* von den t-Werten liegen.

Die kritische Grenze lesen wir daher bei einem 1 %igen Signifikanzniveau an der Stelle $t_{(1-0{,}01;31)}$ ab. Aus der t-Tabelle entnehmen wir für $df = 30$ (da die Werte für eine Verteilung mit $df = 31$ in der Tabelle nicht vorliegen) in der Spalte $(1-\alpha) = 0{,}99$ den Wert 2,457. Die Wahrscheinlichkeit bei Gültigkeit der H_0 t-Werte zu erhalten, die größer als 2,46 sind, ist kleiner als 1 %. Solche Abweichungen sind bei Gültigkeit der Nullhypothese also sehr unwahrscheinlich.

Wir lehnen die Nullhypothese daher ab, wenn die Prüfgröße größer als 2,46 ist. Die Nullhypothese wird angenommen, wenn die Prüfgröße kleiner als 2,46 ist.

$$t > 2{,}46 \quad \longrightarrow \quad H_0 \text{ ablehnen}$$
$$t \leq 2{,}46 \quad \longrightarrow \quad H_0 \textbf{ nicht} \text{ ablehnen}$$

4. Prüfgröße berechnen und Entscheidung über die Nullhypothese treffen

Beim Einsetzen der Werte resultiert:

$$t = \frac{\bar{x}_d}{\frac{\hat{\sigma}_d}{\sqrt{n}}} = \frac{13}{\frac{6}{\sqrt{32}}} = 12{,}26\,.$$

Da 12,26 größer als 2,46 ist, kann die Nullhypothese verworfen werden. Die Verbesserung in der Statistikklausur durch den Besuch des Statistikkurses um durchschnittlich 13 Punkte ist hochsignifikant.

12.4 χ^2-Test auf Unabhängigkeit

Ein Test für die Unabhängigkeit von zwei diskreten Merkmalen ist der χ^2-Unabhängigkeitstest.[3] Die Prüfgröße dieses Tests ist das Maß χ^2, das wir in Kapitel 7.3.1 eingeführt haben. Mit dem χ^2-Unabhängigkeitstest wird geprüft, ob zwei Merkmale in der Grundgesamtheit unabhängig sind.

Untersucht werden soll der Zusammenhang zwischen dem Geschlecht und der Einstellung zum Schwangerschaftsabbruch. Im ALLBUS 1996 wurden zum Schwangerschaftsabbruch eine Reihe von Fragen gestellt. Gefragt wurde unter anderem, ob es gesetzlich möglich oder nicht möglich sein sollte, dass eine Frau einen Schwangerschaftsabbruch vornehmen lässt, unabhängig davon, welche Gründe sie hat. Das folgende Beispiel beschränkt sich auf Westdeutschland.

1. Null- und Alternativhypothese formulieren, Signifikanzniveau festlegen

Wie bei jedem Test werden zunächst Hypothesen über die Grundgesamtheit – im Beispiel also Westdeutschland – aufgestellt. Wir vermuten, dass Frauen eher als Männer der Meinung sind, ein Schwangerschaftsabbruch solle legal sein. Dies ist die Alternativhypothese H_A, die einen Zusammenhang zwischen den beiden Merkmalen Geschlecht und Einstellung zum

3 Der hier behandelte χ^2-Test ist ein Test für unabhängige Stichproben. Bei abhängigen Stichproben müssen andere Verfahren angewendet werden.

Schwangerschaftsabbruch postuliert. Die Behauptung, dass kein Zusammenhang zwischen dem Geschlecht und der Einstellung zum Schwangerschaftsabbruch existiert, entspricht dem Inhalt der Nullhypothese. Die H_0 lautet also, dass ein Zusammenhang zwischen beiden Merkmalen in der Grundgesamtheit nicht existiert.

H_0 : Die Merkmale sind statistisch unabhängig.
H_A : Die Merkmale sind statistisch abhängig.

Die Nullhypothese soll auf einem Signifikanzniveau von 5 % getestet werden.

2. Prüfgröße und Verteilung der Prüfgröße bestimmen

Wir haben in Kapitel 7.1 gesehen, dass bei statistischer Unabhängigkeit die prozentuale Verteilung des abhängigen Merkmals für jede Ausprägung des unabhängigen Merkmals identisch ist. Die bei statistischer Unabhängigkeit **erwarteten Häufigkeiten** $f_{e(ij)}$ lassen sich nach Gleichung 7.1 (S. 144) ermitteln:

$$f_{e(ij)} = \frac{\text{Zeilensumme} \cdot \text{Spaltensumme}}{n} \,. \tag{12.16}$$

Die Tabelle, die die erwarteten Häufigkeiten beinhaltet, wird als Indifferenztabelle bezeichnet. Bei einem χ^2-Test werden die erwarteten Häufigkeiten mit den in der Stichprobe **beobachteten Häufigkeiten** $f_{b(ij)}$ verglichen.

Je größer die Differenz zwischen beobachteten und erwarteten Werten $f_{b(ij)} - f_{e(ij)}$, umso stärker weichen die beobachteten Häufigkeiten vom Modell statistischer Unabhängigkeit ab. Da die Summe der einfachen Differenzen für alle Zellen null ist, werden die Differenzen quadriert. Größere Abweichungen werden hierdurch stärker gewichtet als kleine. Die quadrierte Abweichung in einer Zelle $(f_{b(ij)} - f_{e(ij)})^2$ wird außerdem durch die erwartete Häufigkeit $f_{e(ij)}$ dividiert, da eine bestimmte Abweichung bei einer kleinen erwarteten Häufigkeit stärker ins Gewicht fällt als bei einer großen.

Zur Berechnung von χ^2 (Prüfgröße) werden die quadrierten und relativierten Abweichungen aller Zellen addiert (vgl. Gleichung 7.8).

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{m} \frac{(f_{b(ij)} - f_{e(ij)})^2}{f_{e(ij)}} \tag{12.17}$$

χ^2 nimmt den Wert null an, wenn beobachtete und erwartete Häufigkeiten in allen Zellen übereinstimmen. Es wird umso größer, je weiter beobachtete und erwartete Häufigkeiten auseinanderfallen.

Auch wenn beide Merkmale in der Grundgesamtheit statistisch unabhängig sind, kann man – aufgrund zufälliger Abweichungen der Stichprobe von der Grundgesamtheit – nicht davon ausgehen, dass der für eine Stichprobe ermittelte χ^2-Wert exakt null ist. Die χ^2-Verteilung gibt die Wahrscheinlichkeit von χ^2-Werten bei gegebenen *df* in Stichproben an, wenn in der Grundgesamtheit die Nullhypothese gilt.

χ^2-Verteilung

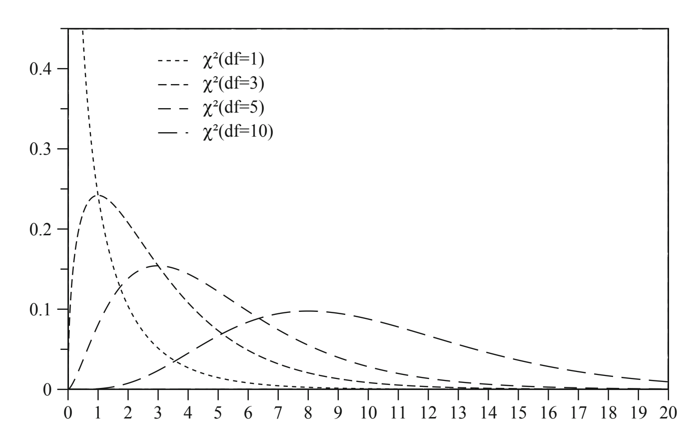
Die **χ^2-Verteilung** ist im Gegensatz zur z- und t-Verteilung keine symmetrische Verteilung. Ihre Form und Lage ist abhängig von der **Zahl der Freiheitsgrade** (*df*). Wie man in Abbildung 12.5 sieht, verschiebt sich die Verteilung mit zunehmenden Freiheitsgraden auf der x-Achse weiter nach rechts. Der Mittelwert der χ^2-Verteilung entspricht der Zahl der Freiheitsgrade (*df*), während die Varianz sich aus $2 \cdot df$ ergibt.

Die Freiheitsgrade werden bei einem χ^2-Unabhängigkeitstest aus der Zahl der Spalten und der Zahl der Zeilen berechnet.

$$df = (\text{Zahl der Zeilen} - 1) \cdot (\text{Zahl der Spalten} - 1) \tag{12.18}$$

Der χ^2-Wert einer Tabelle mit 5 Zeilen und 4 Spalten hat also $(5 - 1) \cdot (4 - 1) = 12$ Freiheitsgrade. Die Freiheitsgrade geben an, wie viele der Zelleninhalte einer Tabelle (bei gegebener Randverteilung) frei variieren können, bevor die anderen Zelleninhalte festgelegt sind. In einer 2×2-Tabelle kann $(2 - 1) \cdot (2 - 1) = 1$ Zelleninhalt frei variieren. Legt man

Abbildung 12.5: χ^2-Verteilung für verschiedene Freiheitsgrade



einen Zelleninhalt fest, dann können bei einer 2×2-Tabelle alle anderen Zellhäufigkeiten als Differenz zu den Randhäufigkeiten ermittelt werden.

Die Prüfgröße ist allerdings nur dann χ^2-verteilt, wenn die erwarteten Häufigkeiten $f_{e(ij)}$ in den Zellen groß genug sind. Als Faustregel wird angegeben, dass die **erwartete Häufigkeit $f_{e(ij)}$ in jeder Zelle größer als fünf** ist. Ist dies nicht der Fall, dann können – sofern dies sinnvoll erscheint – Kategorien zusammengefasst werden, bevor der Test durchgeführt wird. Ansonsten sollte ein Test für kleine Zellbesetzungen verwendet werden (vgl. Agresti 1996, S. 39–45).

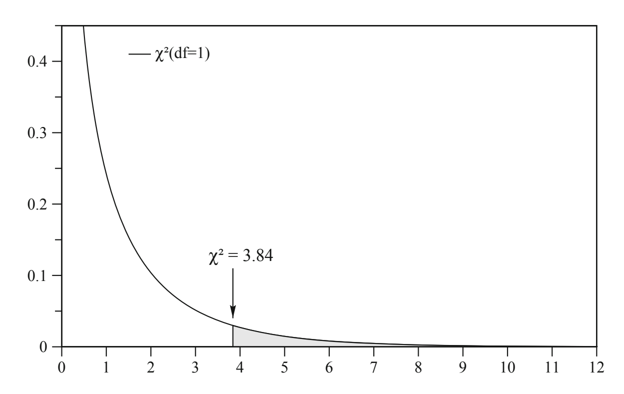
3. Ablehnungsbereich der Nullhypothese kennzeichnen

Bei einem χ^2-Test ist die Frage, wie wahrscheinlich der beobachtete oder ein noch größerer χ^2-Wert bei Gültigkeit der Nullhypothese ist. Der Ablehnungsbereich befindet sich immer am rechten Ende der Verteilung. Gesucht wird also der χ^2-Wert, der am rechten Rand eine Fläche der Größe α abschneidet. Links vom gesuchten Wert liegt $1 - \alpha$ der Fläche.

Der χ^2-Tabelle entnehmen wir für ein Signifikanzniveau von $\alpha = 0,05$ bei einer Verteilung mit einem Freiheitsgrad in der Spalte „0,95" $(1 - \alpha)$ den χ^2-Wert 3,84. Gilt in der Grundgesamtheit die Nullhypothese, dann sind

χ^2-Werte größer als 3,84 unwahrscheinlicher als 5 %. Die Nullhypothese lehnen wir bei einem Signifikanzniveau von 5 % und einer Verteilung mit einem Freiheitsgrad also ab, wenn in der Stichprobe ein χ^2-Wert größer als 3,84 ermittelt wird. Wir lehnen die Nullhypothese nicht ab, wenn der χ^2-Wert der Stichprobe kleiner oder gleich 3,84 ist. Der Ablehnungsbereich ist in Abbildung 12.6 grau schraffiert.

Abbildung 12.6: Ablehnungsbereich in einer χ^2-Verteilung mit $df = 1$ bei einem Signifikanzniveau von 5 %



$$\chi^2 > 3,84 \quad \longrightarrow \quad H_0 \text{ ablehnen}$$
$$\chi^2 \leq 3,84 \quad \longrightarrow \quad H_0 \textbf{ nicht} \text{ ablehnen}$$

4. Prüfgröße berechnen und Entscheidung über die Nullhypothese treffen

Tabelle 12.2 gibt den Zusammenhang zwischen dem Geschlecht und der Einstellung zum Schwangerschaftsabbruch bei den westdeutschen Befragten wieder. 35,6 % (377 von 1058) der Männer und 37,2 % (397 von 1067) der Frauen geben an, dass der Schwangerschaftsabbruch unabhängig von den Motiven einer Frau legal sein sollte. In ihrer Einstellung zum Schwangerschaftsabbruch unterscheiden sich Männer und Frauen in der Stichprobe also nur geringfügig. Mit dem χ^2-Test wird nun geprüft, ob beide Merkmale in der Grundgesamtheit unabhängig sind.

Tabelle 12.2: Kontingenztabelle – Einstellung zum Schwangerschaftsabbruch und Geschlecht

	Männer	Frauen	Summe
sollte legal sein	377	397	774
sollte nicht legal sein	681	670	1351
Summe	1058	1067	2125

Quelle: ALLBUS 1996, westdeutsche Befragte

In Tabelle 12.3 sind die erwarteten Häufigkeiten wiedergegeben, die sich nach Gleichung 12.16 (S. 299) berechnen. Bei statistischer Unabhängigkeit beider Merkmale müssten 36,4 % (385,4 von 1058) der Frauen und 36,4 % (388,6 von 1067) der Männer die Meinung vertreten, ein Schwangerschaftsabbruch solle ohne Angaben von Gründen legal sein.

Tabelle 12.3: Indifferenztabelle – Einstellung zum Schwangerschaftsabbruch und Geschlecht

	Männer	Frauen	Summe
sollte legal sein	385,4	388,6	774
sollte nicht legal sein	672,6	678,4	1351
Summe	1058	1067	2125

Quelle: ALLBUS 1996, westdeutsche Befragte

In unserem Beispiel weichen die in der Stichprobe beobachteten Häufigkeiten nicht sehr weit von den erwarteten Häufigkeiten ab. χ^2 berechnet sich nach Gleichung 12.17:

$$\chi^2 = \frac{(377 - 385{,}4)^2}{385{,}4} + \frac{(397 - 388{,}6)^2}{388{,}6} + \frac{(681 - 672{,}6)^2}{672{,}6} + \frac{(670 - 678{,}4)^2}{678{,}4}$$
$$= 0{,}568\,.$$

Da der empirisch ermittelte χ^2-Wert von 0,568 kleiner als der kritische χ^2-Wert von 3,84 ist, wird die Nullhypothese *nicht* verworfen. Bei Gültigkeit der Nullhypothese (und $df = 1$) ist die Wahrscheinlichkeit sehr groß, in einer Stichprobe einen χ^2-Wert von 0,568 oder größer zu erhalten (vgl. Abbildung 12.6). Stata ermittelt einen p-Wert von $0,45 = 45\,\%$. Die

Wahrscheinlichkeit, bei Gültigkeit der H_0 (und $df = 1$) in einer Stichprobe einen χ^2-Wert größer als 0,568 zu erhalten, beträgt 45 %.

Zu beachten ist, dass mit dem χ^2-Test zwar die statistische Unabhängigkeit von zwei Merkmalen in der Grundgesamtheit überprüft werden kann, nicht aber, wie stark der Zusammenhang ist. Wie wir wissen (vgl. Kapitel 7) ist die Größe des χ^2-Werts – bei gleicher prozentualer Verteilung – direkt proportional zur Stichprobengröße.

Ein angemessenes Zusammenhangsmaß für zwei nominalskalierte Merkmale ist z. B. Cramérs V (vgl. Gleichung 7.11, S. 151).

$$\text{Cramérs V} = \sqrt{\frac{0,568}{2125 \cdot (2-1)}} = 0,016$$

Die Stärke des Zusammenhangs zwischen dem Geschlecht und der Einstellung zum Schwangerschaftsabbruch ist nahe null.

An diesem Beispiel lässt sich die Bedeutung des Stichprobenumfangs für die Signifikanz veranschaulichen: Bei einer Verzehnfachung der *Zellhäufigkeiten* von Tabelle 12.2 resultiert ein zehnfach größerer χ^2-Wert, nämlich 5, 68, bei gleicher Zahl der Freiheitsgrade. Dieser Wert wäre auf dem 5 %-Niveau statistisch signifikant, da er größer als 3,84 ist. Da sich die Proportionen nicht geändert haben, beträgt Cramérs V auch hier 0, 016. Der Unterschied zwischen Männern und Frauen in der Einstellung zum Schwangerschaftsabbruch wäre also immer noch inhaltlich nicht bedeutsam.

Zusammenfassung

In diesem Kapitel wurden einführend grundlegende Testverfahren vorgestellt. Auch für andere Stichprobenkennwerte – wie Zusammenhangsmaße oder Regressionskoeffizienten – existieren statistische Tests. Die Prüfgröße gibt häufig die Abweichung des Punktschätzers vom Parameter der H_0 in Standardfehlern an, wie wir gesehen haben:

$$\text{Prüfgröße} = \frac{\text{Punktschätzer} - \text{Parameter bei Gültigkeit der} H_0}{\text{Standardfehler des Punktschätzers}} .$$

Zur Illustration soll auch hier wieder auf die Regression aus Kapitel 8 zurückgegriffen werden. Der durch die Stichprobe ($n = 2062$) errechnete Regressionskoeffizient der Lesefähigkeit b beträgt 0,84. b ist der Punktschätzer für den Regressionskoeffizient der Grundgesamtheit β. Getestet werden soll, ob die Lesekenntnisse zur Prognose der Mathematikkenntnisse geeignet sind ($H_A : \beta \neq 0$). Die Nullhypothese postuliert hier, dass der Regressionskoeffizient der Grundgesamtheit β gleich null ist, $H_0 : \beta = 0$. Stata gibt den Standardfehler des Regressionskoeffizienten mit $\hat{\sigma}_b = 0{,}01$ an. Die Prüfgröße berechnet sich nach

$$t = \frac{b - \beta}{\hat{\sigma}_b} = \frac{0,84 - 0}{0,01} = 84\,. \tag{12.19}$$

Diese Prüfgröße ist mit $df = n - 2 = 2062 - 2$ t-verteilt. Bei $df = 2060$ kann die z-Verteilung herangezogen werden. Bei einer 5 %igen Irrtumswahrscheinlichkeit (zweiseitige Fragestellung) wird die Hypothese abgelehnt, wenn $z > |1{,}96|$. Da 84 größer als 1,96 ist, lehnen wir die H_0 ab. Das Konfidenzintervall wurde in Kapitel 11 (S. 270) berechnet. Es beinhaltet nicht den Wert der Nullhypothese.

Aufgaben zu Hypothesenprüfung

1. Wozu benötigt man Testverfahren?
2. Ein Bekannter von Ihnen stellt die Behauptung auf, dass Arbeitslose in Ostdeutschland durchschnittlich nicht länger als ein Jahr (maximal 52 Wochen) arbeitslos seien. Aufgrund der Arbeitsmarktsituation in Ostdeutschland vermuten Sie jedoch, dass die durchschnittliche Dauer der Arbeitslosigkeit über einem Jahr liegt. Ihre Behauptung stellt die Alternativhypothese dar (mehr als 52 Wochen).
   Im ALLBUS 1998 wurde die Dauer der Arbeitslosigkeit in Wochen erfasst. Durchschnittlich waren die 144 ostdeutschen Befragten, die Arbeitslosigkeit angaben, seit 67,7 Wochen ($= \bar{x}$) arbeitslos. Die Standardabweichung der Grundgesamtheit wird durch die Stichprobe geschätzt und beträgt $\hat{\sigma} = 61,5$ Wochen.
   a) Prüfen Sie auf Basis der ALLBUS-Daten auf einem Signifikanzniveau von 5 %, ob die durchschnittliche Dauer der Arbeitslosigkeit nicht mehr als ein Jahr, also 52 Wochen, beträgt.
   b) Berechnen Sie ein 95%iges Konfidenzintervall und interpretieren Sie dieses.
3. Sie möchten für Westdeutschland untersuchen, ob sich Frauen und Männer ideologisch unterscheiden. Als Indikator für die politische Ideologie ziehen Sie die Links-Rechts-Skala heran, die im ALLBUS 1998 enthalten ist. Auf einer zehnstufigen Skala konnten die Befragten sich von 1 (ganz links) bis 10 (ganz rechts) einordnen. Wir unterstellen, dass die Links-Rechts-Skala intervallskaliert ist, zwischen den Skalenpunkten also gleiche Abstände bestehen.
   Für die 1.083 Frauen (n_1) wurde ein durchschnittlicher Skalenwert von $\bar{x}_1 = 5,06$ bei einer Standardabweichung $\hat{\sigma}_1 = 1,58$ Skalenpunkten ermittelt; für die 987 Männer (n_2) wurde ein durchschnittlicher Skalenwert von $\bar{x}_2 = 5,25$ und eine Standardabweichung von $\hat{\sigma}_2 = 1,74$ Skalenpunkten berechnet.
   a) Formulieren Sie die Null- und Alternativhypothese. Prüfen Sie mit einem z-Test für Mittelwertunterschiede, ob der Unterschied in der ideologischen Selbsteinstufung von Männern und Frauen ($\bar{x}_1 - \bar{x}_2 = 0,19$) statistisch signifikant ist. Legen Sie ein Signifikanzniveau von 1 % zugrunde.
   b) Berechnen Sie außerdem das Konfidenzintervall für die Differenz der Mittelwerte für eine Vertrauenswahrscheinlichkeit von 99 %. In

welchem Bereich liegt der „wahre“ Unterschied zwischen Männern und Frauen?

c) Berechnen Sie η!

4. Bitte prüfen Sie mit Hilfe des χ^2-Tests, ob der auf Seite 175 dargestellte Zusammenhang zwischen der Konfessionszugehörigkeit und der Wahlabsicht auf einem Signifikanzniveau von $\alpha = 0{,}05$ signifikant ist.
5. Sie haben einen Signifikanztest durchgeführt. Ein Statistik-Programm gibt einen p-Wert von 0,02 an.
   a) Welche Entscheidung treffen Sie bei einem Signifikanzniveau von $\alpha = 0{,}05$? Wenn die Entscheidung falsch ist – welchen Fehler begehen Sie?
   b) Welche Entscheidung treffen Sie bei einem Signifikanzniveau von $\alpha = 0{,}01$? Wenn die Entscheidung falsch ist – welchen Fehler begehen Sie?

Anhang A

Tabellen zur Berechnung der Fläche unter den Wahrscheinlichkeitsverteilungen

z-Verteilung

z-Wert	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
-2,9.	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8.	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7.	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6.	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5.	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4.	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3.	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2.	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1.	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0.	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9.	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8.	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7.	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6.	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5.	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4.	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3.	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2.	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1.	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0.	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9.	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8.	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7.	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6.	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5.	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4.	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3.	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2.	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1.	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
-0,0.	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2,

(Fortsetzung z-Verteilung)

z-Wert	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
0,0.	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1.	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2.	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3.	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4.	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5.	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6.	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7.	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8.	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9.	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0.	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1.	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2.	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3.	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4.	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5.	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6.	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7.	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8.	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9.	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0.	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1.	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2.	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3.	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4.	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5.	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6.	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7.	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8.	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9.	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Lesehilfe: Gesucht sei der Flächenanteil, der zwischen $-\infty$ und dem Wert $z = 1{,}96$ liegt. In der Spalte „z-Wert" am linken Rand der Tabelle sucht man zunächst die Zeile mit dem Wert „1,9." (Der Punkt steht für alle Ziffern von 0 bis 9). Dann sucht man in dieser Zeile die Spalte mit dem Wert „.6". Dort kann man der Tabelle den Flächenanteil „0,9750" entnehmen, also 97,5%.

Gesucht sei ferner der z-Wert, der linksseitig 2,5% der Fläche abschneidet. Der Wert, der nun innerhalb der Tabelle zu suchen ist, beträgt „0,0250". Man findet ihn in der Zeile „$-1{,}9$." und der Spalte „.6". Also teilt der z-Wert $-1{,}96$ linksseitig 2,5% der Fläche ab.

t-Verteilung

df	Fläche $(1-\alpha)$ 0,65	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656
2	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
30	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
40	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704
50	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678
60	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660
70	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,994	2,381	2,648
80	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639
90	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,368	2,632
100	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626
150	0,386	0,526	0,676	0,844	1,040	1,287	1,655	1,976	2,351	2,609
200	0,386	0,525	0,676	0,843	1,039	1,286	1,653	1,972	2,345	2,601
500	0,386	0,525	0,675	0,842	1,038	1,283	1,648	1,965	2,334	2,586
1000	0,385	0,525	0,675	0,842	1,037	1,282	1,646	1,962	2,330	2,581
z-Wert	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576

Lesehilfe: Welcher t-Wert schneidet linksseitig von der t-Verteilung mit 17 Freiheitsgraden 95% der Fläche ab? In der Spalte „df" am linken Rand der Tabelle sucht man die Zeile mit dem Wert „17" und dann in dieser Zeile die Spalte mit dem Wert „0,95". Hier findet man den t-Wert 1,740. Da die t-Verteilung mit zunehmender Anzahl an Freiheitsgraden in eine Normalverteilung übergeht, ist am Fuß der Tabelle der entsprechende z-Wert wiedergegeben (vgl. Abbildung 11.5 auf Seite 265).

χ^2-Verteilung

	Fläche (1-α)								
df	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995
1	1,07	1,32	1,64	2,07	2,71	3,84	5,02	6,63	7,88
2	2,41	2,77	3,22	3,79	4,61	5,99	7,38	9,21	10,60
3	3,66	4,11	4,64	5,32	6,25	7,81	9,35	11,34	12,84
4	4,88	5,39	5,99	6,74	7,78	9,49	11,14	13,28	14,86
5	6,06	6,63	7,29	8,12	9,24	11,07	12,83	15,09	16,75
6	7,23	7,84	8,56	9,45	10,64	12,59	14,45	16,81	18,55
7	8,38	9,04	9,80	10,75	12,02	14,07	16,01	18,48	20,28
8	9,52	10,22	11,03	12,03	13,36	15,51	17,53	20,09	21,95
9	10,66	11,39	12,24	13,29	14,68	16,92	19,02	21,67	23,59
10	11,78	12,55	13,44	14,53	15,99	18,31	20,48	23,21	25,19
11	12,90	13,70	14,63	15,77	17,28	19,68	21,92	24,73	26,76
12	14,01	14,85	15,81	16,99	18,55	21,03	23,34	26,22	28,30
13	15,12	15,98	16,98	18,20	19,81	22,36	24,74	27,69	29,82
14	16,22	17,12	18,15	19,41	21,06	23,68	26,12	29,14	31,32
15	17,32	18,25	19,31	20,60	22,31	25,00	27,49	30,58	32,80
16	18,42	19,37	20,47	21,79	23,54	26,30	28,85	32,00	34,27
17	19,51	20,49	21,61	22,98	24,77	27,59	30,19	33,41	35,72
18	20,60	21,60	22,76	24,16	25,99	28,87	31,53	34,81	37,16
19	21,69	22,72	23,90	25,33	27,20	30,14	32,85	36,19	38,58
20	22,77	23,83	25,04	26,50	28,41	31,41	34,17	37,57	40,00
21	23,86	24,93	26,17	27,66	29,62	32,67	35,48	38,93	41,40
22	24,94	26,04	27,30	28,82	30,81	33,92	36,78	40,29	42,80
23	26,02	27,14	28,43	29,98	32,01	35,17	38,08	41,64	44,18
24	27,10	28,24	29,55	31,13	33,20	36,42	39,36	42,98	45,56
25	28,17	29,34	30,68	32,28	34,38	37,65	40,65	44,31	46,93
30	33,53	34,80	36,25	37,99	40,26	43,77	46,98	50,89	53,67
40	44,16	45,62	47,27	49,24	51,81	55,76	59,34	63,69	66,77
50	54,72	56,33	58,16	60,35	63,17	67,50	71,42	76,15	79,49
60	65,23	66,98	68,97	71,34	74,40	79,08	83,30	88,38	91,95
70	75,69	77,58	79,71	82,26	85,53	90,53	95,02	100,43	104,21
80	86,12	88,13	90,41	93,11	96,58	101,88	106,63	112,33	116,32
90	96,52	98,65	101,05	103,90	107,57	113,15	118,14	124,12	128,30
100	106,91	109,14	111,67	114,66	118,50	124,34	129,56	135,81	140,17
150	158,58	161,29	164,35	167,96	172,58	179,58	185,80	193,21	198,36
200	209,99	213,10	216,61	220,74	226,02	233,99	241,06	249,45	255,26
500	516,09	520,95	526,40	532,80	540,93	553,13	563,85	576,49	585,21
z-Wert	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576

Lesehilfe: Die Vorgehensweise entspricht der der t-Verteilung. Der entsprechende Wert der z-Verteilung am Fuß der Tabelle ergibt sich erst nach Abzug der Freiheitsgrade und Division durch $\sqrt{2 \cdot df}$, da die χ^2-Verteilung mit zunehmender Anzahl an Freiheitsgraden in eine Normalverteilung mit den Parametern $\mu = df$ und $\sigma^2 = 2 \cdot df$ übergeht. Da die Annäherung sehr viel langsamer geschieht als bei der t-Verteilung, stimmen die Werte noch nicht sehr genau überein.

Anhang B

Lösungen der Übungsaufgaben

Forschungsdesigns

1. Individualdaten beziehen sich meist auf Personen, Aggregatdaten auf Kollektive, wobei Aggregatdaten auf der Zusammenfassung von Meßwerten der Mitglieder der Kollektive beruhen.
2. Schließt man aus einem bestehenden Zusammenhang zwischen Arbeitslosenquote und den Stimmanteilen der Republikaner auf Wahlkreisebene, dass Arbeitslose verstärkt Republikaner wählen, so begeht man einen *ökologischen Fehlschluss*. Aufgrund der Aggregatdaten kann man nicht wissen, ob tatsächlich Arbeitslose die Republikaner gewählt haben oder z. B. vor allem Nicht-Arbeitslose in Wahlkreisen mit einer hohen Arbeitslosenquote zur Wahl der Republikaner tendieren.
3. Mit Trenddaten lassen sich Hypothesen über Veränderungen im Aggregat (Nettoveränderungen) überprüfen. *Panelstudien* sind aufwändiger zu erheben und bergen eigene Probleme (Panelmortalität, Repräsentativität etc.). Kausale Hypothesen sind besser prüfbar, weil Informationen über die zeitliche Reihenfolge von Merkmalen vorhanden sind. Individuelle Veränderungen können im Zeitverlauf untersucht werden.
4. Die Zustimmung zu einer traditionellen Arbeitsteilung zwischen Männern und Frauen ging bei westdeutschen Befragten kontinuierlich zurück und sank zwischen 1982 und 2004 um 28 Prozentpunkte bzw. 40 %.
   Da hier die Entwicklung eines Indikators über mehrere Zeitpunkte betrachtet wird, handelt es sich um Längsschnittdaten. Angegeben sind Veränderungen im Aggregat (Nettoveränderungen). Aus diesem Grund handelt es sich um eine Trendstudie. Individuelle Veränderungen lassen sich mit dem ALLBUS auch nicht feststellen, da für jede Erhebung eine neue Stichprobe erhoben wird. Beschränkt man sich bei der Auswertung auf einen ALLBUS, so handelt es sich um eine Querschnittanalyse.
5. Mit einem Experiment, wenn die Teilnehmer des Experiments den Versuchsbedingungen (Gruppen) zufällig zugeteilt wurden.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2,

6. Bei den Volkszählungsdaten von 1987 handelt es sich eigentlich um Individualdaten. Da die Volkszählungsergebnisse jedoch nur in aggregierter Form, z. B. auf Gemeinde- oder Kreisebene, von den Statistischen Ämtern zugänglich gemacht werden, können wir sie lediglich als Aggregatdaten verwenden.

Messen

1. Messen ist die *strukturtreue* Zuordnung von Zahlen zu Objekten.
2. Nominalskala (Gleich/Ungleich) – Ordinalskala (zusätzlich: Ordnung) – Intervallskala (zusätzlich: gleiche Abstände) – Ratioskala (zusätzlich: gleiche Verhältnisse) – Absolutskala (zusätzlich: natürliche Maßeinheit).
3. nominal – ordinal – ratio – ratio – nominal – ordinal – ordinal – nominal – intervall.
4. Antwort (c) ist falsch. Bei Ordinalskalen kann man – im Gegensatz zu Intervallskalen – nicht davon ausgehen, dass die Abstände zwischen den einzelnen Skalenwerten gleich groß sind.
5. Die Größe des Messfehlers kann berechnet werden.
6. Vgl. dazu ausführlich Kapitel 3.3.
7. Der *Reproduzierbarkeitskoeffizient* und das Ergebnis der *Item-Analyse* sagen etwas über die Güte der Guttman- bzw. der Likert-Skala aus.
8. Das Messinstrument scheint sowohl reliabel als auch valide zu sein. Auf die Reliabilität deutet die Tatsache hin, dass verschiedene Messungen zu stabilen Ergebnissen führen. Wäre das Messinstrument nicht valide, würde es nicht mit dem bereits bewährten Messinstrument korrelieren.

Erhebungsmethoden

1. a) Diese Frage wird die meisten Befragten überfordern. Zudem ist der Begriff „Kosten“ nicht eindeutig (Kosten des Studierenden oder der Hochschule?).
   b) Hier handelt es sich um eine mehrdimensionale Frage; man könnte beispielsweise durchaus für eine Straffung des Studiums und gegen die Einführung von Studiengebühren sein. Aus diesem Grunde sind hier zwei Fragen angebracht. Eventuell könnte man eine „weiß nicht“-Kategorie einführen.

c) Die Antwortkategorien überlappen sich („mehr als einmal wöchentlich“ beinhaltet „zwei- bis dreimal wöchentlich“ und „täglich“)
d) Bei dieser Frage sind die Antwortkategorien alles andere als erschöpfend. Wichtige Antwortmöglichkeiten (z. B. Interesse am Fach) fehlen. Hier bietet sich eine Hybridfrage an.
e) Aufgrund der doppelten Verneinung ist kaum klar, welche inhaltliche Bedeutung mit der Zustimmung bzw. Ablehnung der Frage verbunden ist. Die „weiß nicht“-Kategorie sollte nicht die Mittelposition bei den Antwortkategorien einnehmen. Kreuzt ein Befragter hier „weiß nicht“ an, so kann man sich nicht sicher sein, ob er tatsächlich keine Meinung hat oder z. B. der Meinung ist, dass der ASTA teilweise die Meinung der Studierenden vertritt.
f) Diese Frage ist eindeutig zu lang und vor allem unnötig kompliziert formuliert.
g) Hier stimmt alles.

Bei der vorliegenden Anordnung der Fragen könnte es zudem passieren, dass die Frage nach der Höhe der Ausbildungskosten die Beantwortung der darauffolgenden Frage – Einführung von Studiengebühren – beeinflusst.

2. Geschlossene Fragen bieten sich an, wenn die Antwortmöglichkeiten bekannt und/oder begrenzt sind, z. B. beim Geschlecht (Mann – Frau). Offene Fragen bieten sich an, wenn man noch nicht weiß, welche Antworten kommen könnten, wenn die Antwortmöglichkeiten zu zahlreich wären oder wenn man prinzipiell dem Befragten die Gelegenheit geben möchte, ausführlicher oder in seinen eigenen Worten zu antworten.

Tabellen und Graphiken

1. In der Regel werden die Stimmanteile der Parteien auf die gültigen Stimmen prozentuiert, da dieser Anteil über die Sitzverteilung im Parlament entscheidet.
2. Der Anteil der Stimmen für die NSDAP stieg um 103,5 % Prozent ($\frac{37{,}3-18{,}33}{18{,}33}$) bzw. 18,97 Prozentpunkte.
3. Ein Kreis-, Säulen- oder Balkendiagramm, da es sich um ein nominalskaliertes Merkmal handelt.
4. Ostdeutsche Befragte haben insgesamt positivere Erwartungen bezüglich der eigenen wirtschaftlichen Lage in einem Jahr als westdeutsche Befragte (vgl. die Tabelle auf Seite 316).

	Wahlergebnis	(a)	(b)
Wahlberechtigte	42.957.675		=100 %
Abgegebene Stimmen	35.225.758		
Wahlbeteiligung	82 %		
Ungültige Stimmen	254.901		
% Ungült. Stimmen	0,72 %		
Gültige Stimmen	34.970.857	=100 %	
KPD	4.592.090	13,13 %	10,69 %
USPD	11.902	0,03 %	0,03 %
SPD	8.577.738	24,53 %	19,97 %
DDP	1.322.385	3,78 %	3,08 %
Zentrum	4.127.910	11,80 %	9,61 %
BVP	1.059.141	3,03 %	2,47 %
DVP	1.659.774	4,75 %	3,86 %
DNVP	2.458.246	7,03 %	5,72 %
NSDAP	6.409.610	18,33 %	14,92 %
Sonstige	4.752.061	13,59 %	11,06 %

Unter den Befragten, die ihre eigene wirtschaftliche Lage in einem Jahr als wesentlich besser einschätzen, sind Ostdeutsche deutlich überrepräsentiert (36,7 %, im Vergleich zu 32 % ostdeutschen Befragten insgesamt), während Westdeutsche unterrepräsentiert sind (63,3 % zu 68 % insgesamt). Gleiches gilt für die Kategorie „etwas besser“. Allerdings sind Ostdeutsche in der Kategorie „wesentlich schlechter“ ebenfalls deutlich überrepräsentiert, während die westdeutschen Befragten überproportional häufig die Kategorie „etwas schlechter“ angaben.
Von allen Ostdeutschen haben 25,6 % positive Erwartungen an die Entwicklung ihrer wirtschaftlichen Lage, während es bei den Westdeutschen lediglich 16,9 % sind – wenn die Kategorien „wesentlich besser“ und „etwas besser“ zusammengefasst werden.

Lage- und Streuungsparameter

1. Die Verteilung hat zwei Modalwerte: $x_{Mo} = 100$ und $x_{Mo} = 110$; $\tilde{x} = 105$; $\bar{x} = 105{,}1$; $V = 90$; $s^2 = 546{,}09$; $s = 23{,}37$
2. In Land B sind die Einkommens*unterschiede* erheblich geringer ausgeprägt, da die Standardabweichung, also die Streuung der Werte, geringer ist als in Land A.
3. $x_{Mo} = 21$, $x_{Mo} = 22$; $\tilde{x} = 22$; $\bar{x} = 22{,}75$; $V = 10$; $s^2 = 4{,}05$; $s = 2{,}01$

Zeilenprozente *Spaltenprozente* *Totalprozente*	*West*	*Ost*	Summe
wesentlich besser	63,3 1,7 1,1	36,7 2,0 0,7	100,0 1,8 1,8
etwas besser	57,8 15,2 10,3	42,2 23,6 7,5	100,0 17,9 17,9
gleichbleibend	70,6 69,3 47,2	29,4 61,4 19,6	100,0 66,8 66,8
etwas schlechter	71,1 12,8 0,7	28,9 11,0 0,6	100,0 12,2 1,3
wesentlich schlechter	52,3 1,0 0,7	47,7 1,9 0,6	100,0 1,3 1,3
Summe	68,0 100,0 68,0	32,0 100,0 32,0	100,0 100,0 100,0

Die Verteilung hat zwei Modalwerte, nämlich 21 und 22 Jahre. Die Hälfte der Kursteilnehmer hat das 22. Lebensjahr bereits erreicht und im Durchschnitt sind die Teilnehmer 22,75 Jahre alt.
Der älteste und der jüngste Kursteilnehmer liegen 10 Jahre auseinander. Die Streuung der Werte liegt bei 2,01 Jahren.

4. Da der Modalwert größer als das arithmetische Mittel ist, handelt es sich um eine *rechtssteile* (linksschiefe) Verteilung.
5. Angemessen sind in diesem Fall *Modalwert* und *Median*, da es sich bei Klausurnoten um eine *ordinalskalierte Variable* handelt. Dagegen ist die Berechnung des arithmetischen Mittels für Klausurnoten im strengen Sinne nicht zulässig, da die Abstände zwischen den einzelnen Noten nicht gleich sind und damit kein Intervallskalenniveau vorliegt.
6. Das arithmetische Mittel würde bei einer linkssteilen Verteilung von „Ausreißern“ nach oben verzerrt werden. Gibt es also einige sehr hohe Mieten, ist das arithmetische Mittel größer als der Median.

Zusammenhangsmaße

1. • Berechnung der Prozentwerte:

Zeilenprozente *Spaltenprozente*	n. kath.	kath.	Summe
CDU/CSU	44,3 26,4	55,7 43,6	100,0 33,8
SPD	65,5 43,6	34,5 30,1	100,0 37,8
ANDERE	60,0 30,0	40,0 26,3	100,0 28,4
Summe	56,8 100,0	43,2 100,0	100,0

Von den Befragten, die eine Präferenz für CDU/CSU äußerten, sind 55,7 % katholisch und 44,3 % nicht-katholisch. Im Vergleich zu allen Befragten (43,2 % Katholiken), sind Katholiken unter den CDU/CSU-Wählern also überrepräsentiert.
43,6 % aller Katholiken geben an, CDU/CSU wählen zu wollen, während von allen Befragten lediglich 33,8 % eine Präferenz für die Unionsparteien äußern.

- Berechnung von χ^2, C, Cramérs V

$$\chi^2 = \frac{(236 - 302{,}54)^2}{302{,}54} + \frac{(297 - 230{,}46)^2}{230{,}46} + \frac{(390 - 337{,}73)^2}{337{,}73} + \frac{(205 - 257{,}27)^2}{257{,}27} + \frac{(268 - 253{,}73)^2}{253{,}73} + \frac{(179 - 193{,}27)^2}{193{,}27} = 54{,}41$$

$$C = \sqrt{\frac{54{,}41}{54{,}41 + 1575}} = 0{,}183; \quad C_{max} = \sqrt{\frac{2-1}{2}} = 0{,}707$$

$$\text{Cramérs V} = \sqrt{\frac{54{,}41}{1575 \cdot (2-1)}} = 0{,}186$$

Zwischen der Konfession und der Wahlabsicht besteht ein schwacher Zusammenhang.

- Berechnung von λ

Vorhersage der Wahlabsicht durch die Konfession:

$$\lambda = \frac{(533 + 447) - (236 + 268 + 205 + 179)}{(533 + 447)} = \frac{980 - 888}{980} = 0{,}0939 = 9{,}39\,\%$$

Durch die Kenntnis der Konfession lassen sich die Fehler bei Vorhersage der Wahlabsicht um 9,39 % verringern. Auch λ deutet also auf einen schwachen Zusammenhang hin.

2. γ, da es sich um zwei ordinalskalierte Merkmale handelt.

$$\gamma = \frac{1228543 - 932805}{1228543 + 932805} = \frac{295738}{2161348} = 0{,}1368 = 13{,}68\,\%$$

Durch die Kenntnis der Schulbildung der Interviewer lassen sich die Fehler bei Prognose der Schulbildung der Befragten um knapp 14 % verringern. Es existiert also tatsächlich ein schwacher Zusammenhang zwischen der Schulbildung der Interviewer und der Schulbildung der Interviewten.

3. η^2 bzw. η, da die unabhängige Variable (Geschlecht) nominalskaliert und die abhängige (Alter der Befragten) intervallskaliert ist.
Der „Trick“ zur Lösung der Aufgabe besteht darin, die Summe der Abweichungsquadrate aus der Varianz und der Fallzahl zu ermitteln (vgl. Formel 6.8 auf Seite 135):

$$SAQ_{ges} = s^2_{ges} \cdot n_{ges} = 286{,}1653 \cdot 3442 = 984980{,}9626$$

$$SAQ_{kat} = s^2_{Kat_1} \cdot n_{Kat_1} + s^2_{Kat_2} \cdot n_{Kat_2} = (282{,}8915 \cdot 2320) + (293{,}0079 \cdot 1121) = 984770{,}1359$$

$$\eta^2 = \frac{984980{,}9626 - 984770{,}1359}{984980{,}9626} = 0{,}0002 = 0{,}02\,\%$$

$$\eta = \sqrt{0{,}0002} = 0{,}014$$

Ein Zusammenhang zwischen dem Geschlecht des Interviewers und dem Alter des Befragten besteht nicht.

4. Berechnung von Pearsons r:

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	15	1	−6,2	−2,2	13,64	38,44	4,84
2	30	7	8,8	3,8	33,44	77,44	14,44
3	20	2	−1,2	−1,2	1,44	1,44	1,44
4	24	4	2,8	0,8	2,24	7,84	0,64
5	17	2	−4,2	−1,2	5,04	17,64	1,44
	$\bar{x} =$ 21,2	$\bar{y} =$ 3,2			$SAP =$ 55,8	SAQ_x 142,8	SAQ_y 22,8

$$r = \frac{55{,}8}{\sqrt{142{,}8 \cdot 22{,}8}} = 0{,}98$$

Zwischen der Zahl der täglich verzehrten Eis und der Außentemperatur besteht ein fast perfekter Zusammenhang.

5. a) richtig, b) falsch

Lineare Regression

1. Mit einer linearen Regression kann der Einfluss einer (oder mehrerer) unabhängiger Variablen auf eine metrische abhängige Variable bestimmt werden.
2. a) Bestimmung der Regressionsgleichung (vgl. zur Berechnung die Arbeitstabellen auf S. 327 und 328):

$$b = \frac{\text{SAP}}{\text{SAQ}_x} = \frac{-910{,}59}{6136{,}70} = -0{,}1484$$

$$a = \bar{y} - b \cdot \bar{x} = 39{,}21 - (-0{,}1484 \cdot 54{,}99) = 47{,}37$$

Die Regressionsgleichung lautet also:

$$\hat{y}_i = 47{,}37 - 0{,}1484 \cdot x_i$$

Je höher der Prozentanteil der Katholiken in einem Wahlkreis, umso *schlechter* schneidet die SPD ab (negatives Vorzeichen des Regressionskoeffizienten). Es handelt sich also um eine negative

Beziehung. Nimmt der Anteil der Katholiken in einem Wahlkreis um einen Prozentpunkt zu, dann *verlieren* die Sozialdemokraten rund 0,1484 Prozentpunkte. In einem (hypothetischen) Wahlkreis ohne Katholiken würde die SPD 47,37 Prozent der gültigen Stimmen erhalten.

b) Berechnung des Determinationskoeffizienten R^2:

$$R^2 = \frac{\text{Erklärte-SAQ}_y}{\text{Gesamt-SAQ}_y} = \frac{135{,}53}{269{,}20} = 0{,}50$$

Die unterschiedlichen Wahlerfolge der SPD in den rheinland-pfälzischen Wahlkreisen bei der Bundestagswahl 1994 lassen sich zu 50 % durch den Katholikenanteil erklären. (Obwohl dies immer noch ein relativ hohes R^2 ist, liegt der Wert deutlich niedriger als bei Schätzung des CDU-Anteils. Zur Erklärung der Wahlergebnisse der SPD ist der Katholikenanteil also ein schlechterer Prädiktor als zur Erklärung der CDU-Ergebnisse.)

	Korrelations- und Regressionsrechnung						
Wahlkreis	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
Neuwied	55,55	40,96	0,56	1,75	0,98	0,31	3,06
Ahrweiler	81,99	34,17	27,00	−5,04	−135,93	729,00	25,40
Koblenz	73,14	37,93	18,15	−1,28	−23,14	329,42	1,64
Cochem	70,78	32,84	15,79	−6,37	−100,48	249,32	40,58
Kreuznach	32,60	44,01	−22,39	4,80	−107,47	501,31	23,04
Bitburg	91,40	32,72	36,41	−6,49	−236,09	1325,69	42,12
Trier	87,97	39,60	32,98	,39	12,86	1087,68	0,15
Montabaur	50,76	42,21	−4,23	3,01	−12,69	17,89	9,06
Mainz	51,36	36,55	−3,63	−2,66	9,65	13,17	7,08
Worms	32,81	42,42	−22,18	3,21	−71,20	491,95	10,30
Frankenthal	31,98	43,16	−23,01	3,95	−90,09	529,46	15,60
Ludwigshafen	38,01	40,83	−16,98	1,62	−27,51	288,32	2,62
Neustadt-Speyer	45,61	34,59	−9,38	−4,62	43,31	87,98	21,34
Kaiserslautern	34,89	46,70	−20,10	7,49	−150,55	404,01	56,01
Pirmasens	45,98	41,66	−9,01	2,45	−22,07	81,18	6,00
Südpfalz	55,07	36,93	0,08	−2,28	−0,17	0,01	5,20
	$\bar{\mathbf{x}} =$ **54,99**	$\bar{\mathbf{y}} =$ **39,21**			**SAP = −910,59**	$\mathbf{SAQ_x} =$ **6136,70**	$\mathbf{SAQ_y} =$ **269,20**

	Berechnung des Determinationskoeffizienten				
Wahlkreis	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
Neuwied	39,12	1,84	3,38	−0,09	0,01
Ahrweiler	35,19	−1,02	1,04	−4,02	16,16
Koblenz	36,51	1,42	2,02	−2,70	7,29
Cochem	36,86	−4,02	16,16	−2,35	5,52
Kreuznach	42,53	1,48	2,19	3,32	11,02
Bitburg	33,79	−1,07	1,14	−5,42	29,38
Trier	34,30	5,30	28,09	−4,91	24,11
Montabaur	39,83	2,38	5,66	0,62	0,38
Mainz	39,74	−3,19	10,18	0,53	0,28
Worms	42,50	−0,08	0,01	3,29	10,82
Frankenthal	42,62	0,54	0,29	3,41	11,63
Ludwigshafen	41,73	−0,90	0,81	2,52	6,35
Neustadt-Speyer	40,60	−6,01	36,12	1,39	1,93
Kaiserslautern	42,19	4,51	20,34	2,98	8,88
Pirmasens	40,54	1,12	1,25	1,33	1,77
Südpfalz	39,19	−2,26	5,11	−0,02	0,00
			U. $\mathbf{SAQ_y}$ = 133,79		**E. $\mathbf{SAQ_y}$ = 135,53**

3. • Berechnung aus den Einzelmesswerten:

$$r = \frac{\text{SAP}}{\sqrt{\text{SAQ}_x \cdot \text{SAQ}_y}} = \frac{-910{,}59}{\sqrt{6136{,}70 \cdot 269{,}20}} = -0{,}71$$

• Berechnung aus R^2:

$$|r| = \sqrt{0{,}50} = 0{,}71$$

Die Richtung des Zusammenhangs (das Vorzeichen von r!) muss bei der Berechnung aus R^2 dem Regressionskoeffizienten entnommen werden.

Stichprobenverfahren

1. Stichproben sind erheblich schneller und kostengünstiger durchführbar als Vollerhebungen, schwanken allerdings in ihrer Zusammensetzung zufällig und erlauben daher Schlüsse auf die Grundgesamtheitnur mit einer bestimmten Wahrscheinlichkeit.
2. Mehrstufiges Auswahlverfahren

3. • Begriffe:
   Grundgesamtheit = alle Mainzer Studierende
   Auswahlgesamtheit = alle auf der Liste des Studentensekretariates verzeichneten Studierenden
   undercoverage = Studierende, die sich nach Abfassung der Liste immatrikuliert haben
   overcoverage = zwischenzeitlich exmatrikulierte Studierende
   • Vorgehensweise:
   Einfache, systematische Zufallsstichprobe, d. h. zufällige Auswahl des ersten Studierenden, alle weiteren werden in einem bestimmten Intervall ermittelt. Das Stichprobenintervall beträgt $\frac{28734}{1000} = 28{,}734$. Die zu bestimmende Zufallszahl muß also zwischen 1 und 28 liegen (dann erhält man etwas mehr als 1.000 Studierende). Würde die zufällig gezogene erste Zahl 5 lauten, dann würde der 5., der 33., der 61. Studierende usw. in die Stichprobe gelangen.
4. Zufallsgesteuerte Verfahren bieten die Gewähr, dass jedes Element der Grundgesamtheit (genauer: der Auswahlgesamtheit) die gleiche bzw. eine bekannte Wahrscheinlichkeit größer null hat, in die Stichprobe zu gelangen. Erst dadurch werden Schlüsse auf die Grundgesamtheit möglich.

Wahrscheinlichkeitsverteilungen

1. Die z-Tabelle findet sich in Anhang A.

z-Wert	-2,78	-0,10	0,90	1,96
Fläche links	0,0027	0,4602	0,8159	0,9750
Fläche rechts	0,9973	0,5398	0,1841	0,0250

2. Gesucht: Anteil der z-Werte zwischen -2 und 2:

$$\begin{aligned} P(-2 \leq z \leq 2) &= \Phi_2 - \Phi_{-2} \\ &= 0{,}9772 - 0{,}0228 \\ &= 0{,}9544 = 95{,}44\,\% \end{aligned}$$

3. Durch Mittelwert und Varianz.

4. Zunächst müssen hier die beiden x-Werte 20 und 23 z-transformiert werden. Aus der z-Tabelle kann dann die Fläche zwischen den beiden z-transformierten Werten entnommen werden.
   Gesucht: Größe der Fläche, die zwischen 20 und 23 liegt:

$$\begin{aligned} P(20 \leq X \leq 23) &= P\left(\frac{20-20}{4} \leq Z \leq \frac{23-20}{4}\right) \\ &= P(0 \leq Z \leq 0{,}75) \\ &= \Phi_{0{,}75} - \Phi_0 \\ &= 0{,}7734 - 0{,}5 \\ &= 0{,}2734 = 27{,}34\,\% \end{aligned}$$

5. Mit wachsendem Stichprobenumfang n nähert sich die Verteilung von Stichprobenmittelwerten einer Normalverteilung an.
6. Lösungsweg analog zu Aufgabe 4. Da es sich um eine Stichprobenmittelwerteverteilung handelt, muss hier Gleichung 10.21 zur z-Transformation herangezogen werden.
   Gesucht: Prozentsatz der Stichprobenmittelwerte $\bar{x}$, der zwischen 36,9 und 38,9 Jahren liegt:

$$\begin{aligned} P\left(36{,}9 \leq \bar{X} \leq 38{,}9\right) &= P\left(\frac{36{,}9-37{,}9}{0{,}7} \leq Z \leq \frac{38{,}9-37{,}9}{0{,}7}\right) \\ &= P(-1{,}43 \leq Z \leq 1{,}43) \\ &= \Phi_{1{,}43} - \Phi_{-1{,}43} \\ &= 0{,}9236 - 0{,}0764 \\ &= 0{,}8472 = 84{,}72\,\% \end{aligned}$$

7. a) falsch, b) falsch, c) falch, d) richtig

Konfidenzintervalle

1. Konfidenzintervalle sind Bereiche, die den gesuchten Parameter der Grundgesamtheit mit einer gewissen Wahrscheinlichkeit überdecken.
2.
   - Das Konfidenzintervall wird größer.
   - Das Konfidenzintervall wird größer.
   - Das Konfidenzintervall wird kleiner.

3. Die Varianz der Grundgesamtheit ist unbekannt und wird durch die Stichprobendaten geschätzt. Da die Stichprobe sehr groß ist, werden die Grenzen anhand einer z-Tabelle und nicht anhand einer t-Tabelle abgelesen.
Die allgemeine Formel lautet daher:

$$\bar{x} - z_{(1-\frac{\alpha}{2})} \cdot \frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{(1-\frac{\alpha}{2})} \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

Berechnung des zweiseitigen Konfidenzintervalls bei $\alpha = 0{,}05$:

$$1838{,}39 - z_{(1-\frac{0{,}05}{2})} \cdot \frac{1477{,}68}{\sqrt{1474}} \leq \mu \leq 1838{,}39 + z_{(1-\frac{0{,}05}{2})} \cdot \frac{1477{,}68}{\sqrt{1474}}$$
$$1838{,}39 - 1{,}96 \cdot 38{,}49 \leq \mu \leq 1838{,}39 + 1{,}96 \cdot 38{,}49$$
$$1762{,}95 \leq \mu \leq 1913{,}83$$

4. • 99 %iges Konfidenzintervall für den Anteil der CDU/CSU-Wähler:

$$0{,}425 - z_{(1-\frac{0{,}01}{2})} \cdot \sqrt{\frac{0{,}425 \cdot 0{,}575}{1250}} \leq \theta \leq 0{,}425 + z_{(1-\frac{0{,}01}{2})} \cdot \sqrt{\frac{0{,}425 \cdot 0{,}575}{1250}}$$
$$0{,}425 - 2{,}58 \cdot 0{,}014 \leq \theta \leq 0{,}425 + 2{,}58 \cdot 0{,}014$$
$$0{,}389 \leq \theta \leq 0{,}461$$

   • 99 %iges Konfidenzintervall für die PDS:

$$0{,}035 - 2{,}58 \cdot \sqrt{\frac{0{,}035 \cdot 0{,}965}{1250}} \leq \theta \leq 0{,}035 + 2{,}58 \cdot \sqrt{\frac{0{,}035 \cdot 0{,}965}{1250}}$$
$$0{,}035 - 2{,}58 \cdot 0{,}00519 \leq \theta \leq 0{,}035 + 2{,}58 \cdot 0{,}00519$$
$$0{,}02158 \leq \theta \leq 0{,}0484$$

5. Das Konfidenzintervall soll eine Breite von 1 % aufweisen (KIB=0,01); der Prozentsatz der FDP betrug in der Stichprobe der Forschungsgruppe Wahlen p=0,07; α soll 5 % betragen.

$$n = \frac{4 \cdot 1{,}96^2 \cdot 0{,}07 \quad (1 - 0{,}07)}{0{,}01^2} = 10003{,}5$$

Hypothesenprüfung

1. Mit Testverfahren wird anhand zufällig gezogener Stichproben überprüft, inwieweit Hypothesen über eine Grundgesamtheit zutreffend sind.
2. a) Test eines Mittelwertes bei unbekanntem σ.
   - Null- und Alternativhypothese festlegen:
     $H_0: \quad \mu = \mu_0 = 52$ und $H_A: \quad \mu > \mu_0 > 52$ Wochen, $\alpha = 0,05$.
   - Prüfgröße und Verteilung der Prüfgröße bestimmen:
     Die Stichprobe ist hinreichend groß ($n = 144$). Da σ durch die Stichprobe geschätzt werden muss ($\hat{\sigma}$), ist die Prüfgröße mit $df = n - 1 = 143$ Freiheitsgraden t-verteilt.
   - Ablehnungsbereich der Nullhypothese festlegen:
     Weil die Alternativhypothese gerichtet ist und größere Werte postuliert, liegt der gesamte Ablehnungsbereich am rechten Rand der t-Verteilung. Der kritische t-Wert befindet sich daher an der Stelle $t_{(df,1-\alpha)}$. Den kritischen t-Wert entnehmen wir der t-Verteilung mit 150 Freiheitsgraden (da keine Verteilung mit 143 Freiheitsgraden im Anhang tabelliert ist): $t_{(150,1-0,05)} = 1,655$. Ist der Wert der Prüfgröße größer als 1,655, dann wird die Nullhypothese verworfen; ist der Wert kleiner als 1,655, dann wird die Nullhypothese nicht verworfen.
   - Prüfgröße berechnen und über die Nullhypothese entscheiden:

     $$t = \frac{67,6 - 52}{\frac{61,5}{\sqrt{144}}} = 3,044$$

     Da $3,044$ größer als $1,655$ ist, wird die Nullhypothese abgelehnt. Der Unterschied ist statistisch signifikant.

   b) Konfidenzintervall (für μ bei unbekanntem σ) berechnen:

   $$67,7 - 1,96 \cdot \frac{61,5}{\sqrt{144}} \leq \mu \leq 67,7 + 1,96 \cdot \frac{61,5}{\sqrt{144}}$$

   $$57,7 \leq \mu \leq 77,7$$

3. a) Test einer Mittelwertedifferenz bei unabhängigen Stichproben:
   - Null- und Alternativhypothese festlegen:
     $H_0: \quad \mu_1 = \mu_2$ und $H_A: \quad \mu_1 \neq \mu_2, \quad \alpha = 0,01$.

- Prüfgröße und Verteilung der Prüfgröße bestimmen:
  Die Stichproben sind hinreichend groß. Die Prüfgröße ist hier z-verteilt (vgl. die Anmerkungen in Kapitel 12 zu t-Tests).
- Ablehnungsbereich der Nullhypothese bestimmen:
  Für den zweiseitigen Ablehnungsbereich entnimmt man für $\alpha = 0,01$ der z-Tabelle an den Stellen $z_{0,01/2}$ und $z_{1-0,01/2}$ die kritischen Werte $-2,58$ und $2,58$. Die Nullhypothese wird also abgelehnt, wenn in der Stichprobe ein z-Wert kleiner als $-2,58$ oder größer als $2,58$ ermittelt wird.
- Prüfgröße berechnen und über die Nullhypothese entscheiden:

$$\frac{5,06 - 5,25}{\sqrt{\frac{1,58^2}{1083} + \frac{1,74^2}{987}}} = -2,59$$

  Da der in der Stichprobe ermittelte z-Wert $-2,59$ kleiner als der kritische z-Wert $-2,58$ ist (wenn auch sehr knapp), wird die Nullhypothese abgelehnt. Der Unterschied in der ideologischen Einstellung von Männern und Frauen ist also statistisch sehr signifikant.

b) Konfidenzintervall für eine Mittelwertedifferenz:
Die Differenz beträgt in den Stichproben $\bar{x}_1 - \bar{x}_2 = 5,06 - 5,25 = -0,19$; das Konfidenzintervall berechnet sich nach:

$$-0,19 \pm 2,58 \cdot \sqrt{\frac{1,58^2}{1083} + \frac{1,74^2}{987}}$$

Daraus resultiert:

$$-0,3791 \leq \mu_1 - \mu_2 \leq -0,0009$$

Die obere Grenze des 99%igen Konfidenzintervalls ist ganz nah am Wert der Wert der Nullhypothese!

c) Zur Berechnung von η wird die Summe der Abweichungsquadrate für Männer *und* Frauen ($= SAQ_{ges}$) und die Summe der Abweichungsquadrate für Männer und Frauen getrennt benötigt.
Da $\hat{\sigma} = \sqrt{SAQ/n-1}$ ist, ist $SAQ = \hat{\sigma}^2 \cdot (n-1)$:
$SAQ_{ges} = 1,66^2 \cdot 2069 = 5701,3364$
$SAQ_{\text{Frauen}} = 1,58^2 \cdot 1082 = 2701,1048$
$SAQ_{\text{Männer}} = 1,74^2 \cdot 986 = 2985,2136$
$SAQ_{kat} = SAQ_{\text{Frauen}} + SAQ_{\text{Männer}} = 5686,3184$

$$\eta^2 = \frac{SAQ_{ges} - SAQ_{kat}}{SAQ_{ges}} = \frac{5701,3364 - 5686,3184}{5701,3364} = 0,0026$$

$$\eta = \sqrt{\eta} = \sqrt{0,0026} = 0,05$$

Der Zusammenhang zwischen beiden Merkmalen ist zu vernachlässigen!

4. - Null- und Alternativhypothese festlegen:
     H_0: Es gibt keinen Unterschied im Wahlverhalten zwischen Katholiken und Nicht-Katholiken.
     H_A: Die Konfession hat einen Einfluss auf das Wahlverhalten.
     $\alpha = 0{,}05$
   - Prüfgröße und Verteilung der Prüfgröße festlegen:
     Die Prüfgröße χ^2 berechnet sich nach Gleichung 7.8 und ist mit $df = (3-1)(2-1) = 2$ Freiheitsgraden χ^2-verteilt.
   - Ablehnungsbereich der Nullhypothese kennzeichnen:
     Der kritische Wert für ein Signifikanzniveau von 0,05 liegt in einer Verteilung mit zwei Freiheitsgraden bei $\chi^2_{krit} = 5{,}99$ (Anhang A entnehmen!).

$$\begin{aligned} \chi^2 \leq \chi^2_{krit} \quad &\Longrightarrow \quad H_0 \text{ nicht ablehnen} \\ \chi^2 > \chi^2_{krit} \quad &\Longrightarrow \quad H_0 \text{ ablehnen} \end{aligned}$$

   - Berechnung der Prüfgröße χ^2:
     Alle erwarteten Werte sind größer als 5, d. h. der Test darf angewendet werden. Vgl. die Berechnung auf S. 317.
     $\chi^2 = 54{,}41$
   - Entscheidung über die Nullhypothese:
     Da der gemessene χ^2-Wert 54,41 größer als der kritische χ^2-Wert 5,99 ist, wird die Nullhypothese verworfen.

5. a) H_0 ablehnen, da $p < 0,05$. α-Fehler, b) H_0 nicht ablehnen, da $p > 0,01$. β-Fehler

Literaturverzeichnis

Adorno, Theodor W. et al. (1950): *The Authoritarian Personality.* New York/London: Harper und Row.

Agresti, Alan (1996): *An Introduction to Categorical Data Analysis.* New York u. a.: Wiley.

Agresti, Alan und Barbara Finlay (2008): *Statistical Methods for the Social Sciences.* Upper Saddle River, New Jersey: Prentice Hall, 4. Auflage.

Alba, Richard, Peter Schmidt und Martina Wasmer, Hrsg. (2000): *Blickpunkt Gesellschaft 5. Deutsche und Ausländer: Freunde, Fremde oder Feinde?* Opladen/Wiesbaden: Westdeutscher Verlag.

Allerbeck, Klaus R. (1978): *Meßniveau und Analyseverfahren – Das Problem „strittiger Intervallskalen“*, in: *Zeitschrift für Soziologie* 7(3), S. 199–214.

Allison, Paul D. (2002): *Missing Data*, Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks: Sage.

Andreß, Hans-Jürgen, Jaques A. Hagenaars und Steffen Kühnel (1997): *Analyse von Tabellen und kategorialen Daten. Log-lineare Modelle, latente Klassenanalyse, logistische Regression und GSK-Ansatz.* Berlin/Heidelberg: Springer.

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute, Hrsg. (1999): *Stichproben-Verfahren in der Umfrageforschung. Eine Darstellung für die Praxis.* Opladen: Leske + Budrich, Kommentar.

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (2000): *Jahresbericht 2000.* `http://www.adm-ev.de/pdf/Jahresbericht_00.pdf`.

Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (2006): *Jahresbericht 2006.* `http://www.adm-ev.de/pdf/Jahresbericht_06.pdf`.

Babbie, Earl (1997): *The Practice of Social Research.* Belmont, Ca.: Wadsworth, 8. Auflage.

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen,*
DOI 10.1007/978-3-531-91879-2,

Backhaus, Klaus et al. (2003): *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Berlin/Heidelberg: Springer, 11. Auflage.

Barnes, Samuel et al. (1979): *Political Action. Mass Participation in Five Western Democracies*. Beverly Hills: Sage.

Batinic, Bernard et al., Hrsg. (1999): *Online Research. Methoden, Anwendungen und Ergebnisse*. Göttingen: Hogrefe.

Bäcker, Gerhard (2007): *Was heißt hier „geringfügig"? Minijobs als wachsendes Segment prekärer Beschäftigung*, in: Berndt Keller und Hartmut Seifert, Hrsg.: *Atypische Beschäftigung - Flexibilisierung und soziale Risiken*. Berlin: edition sigma, S. 107–126.

Behnke, Joachim und Nathalie Behnke (2006): *Grundlagen der statistischen Datenanalyse*. Wiesbaden: VS Verlag.

Benninghaus, Hans (2005): *Deskriptive Statistik. Eine Einführung für Sozialwissenschaftler.* Wiesbaden: VS Verlag, 10. Auflage.

Bijleveld, Catrien C. J. H. und Leo J. Th. van der Kamp (1998): *Longitudinal Data Analysis. Design, Models and Methods*. London: Sage.

Billiet, Jaak, Achim Koch und Michel Philippens (2007): *Understand and Improving Response Rates*, in: Roger Jowell et al., Hrsg.: *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey.* Thousand Oaks u. a.: Sage, S. 113–135.

Blalock, Hubert M. (1967): *Toward a Theory of Minority-Group-Relations.* New York: John Wiley & Sons.

Bleymüller, Josef, Günther Gehlert und Herbert Gülicher (2004): *Statistik für Wirtschaftswissenschaftler*. München: Vahlen, 14. Auflage.

Blossfeld, Hans-Peter, Katrin Golsch und Götz Rohwer (2007): *Event History Analysis with Stata*. Mahwah, New Jersey und London: Lawrence Erlbaum.

Böker, Fred (1993): *Statistik lernen am PC. Programmbeschreibungen, Übungen und Lernziele zum Statistikprogrammpaket GSTAT*. Göttingen: Vandenhoeck & Ruprecht, 2. Auflage.

Böker, Fred (1998): *Mehr Statistik lernen am PC. Programmbeschreibungen, Übungen und Lernziele zum Statistikprogrammpaket GSTAT2*. Göttingen: Vandenhoeck & Ruprecht, 2. Auflage.

Böltken, Ferdinand (1976): *Auswahlverfahren*. Stuttgart: Teubner.

Bortz, Jürgen (2004): *Statistik für Human- und Sozialwissenschaftler*. Berlin/Heidelberg: Springer, 6. Auflage.

Bortz, Jürgen und Nicola Döring (2006): *Forschungsmethoden und Evaluation*. Berlin/Heidelberg: Springer, 4. Auflage.

Braun, Michael und Peter Ph. Mohler, Hrsg. (1994): *Blickpunkt Gesellschaft 3. Einstellungen und Verhalten der Bundesbürger*. Opladen: Westdeutscher Verlag.

Braun, Michael und Peter Ph. Mohler, Hrsg. (1998): *Blickpunkt Gesellschaft 4. Soziale Ungleichheit in Deutschland*. Opladen: Westdeutscher Verlag.

Brosius, Felix (2006): *SPSS 14*. Bonn: MITP.

Bundesministerium des Innern (2006): *Deutschland beteiligt sich mit einer registergestützten Zählung an der kommenden Volkszählungsrunde der EU 2010/2011. Pressemitteilung vom 29. August 2006*.

Bürklin, Wilhelm und Markus Klein (1998): *Wahlen und Wählerverhalten. Eine Einführung*. Opladen: Leske + Budrich, 2. Auflage.

Campbell, Angus et al. (1980): *The American Voter. Unabridged Edition*. Chicago: Chicago University Press (Midway Reprint).

Campbell, Donald T. und Donald W. Fiske (1959): *Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix*, in: *Psychological Bulletin* 86, S. 81–105.

Campbell, Donald T. und Julian C. Stanley (1966): *Experimental and Quasi-Experimental Designs for Research*. Boston u. a.: Houghton Mifflin.

Carmines, Edward G. und Richard A. Zeller (1979): *Reliability and Validity Assessment*. Beverly Hills: Sage, 2. Auflage.

Chalmers, Alan F. (2007): *Wege der Wissenschaft. Einführung in die Wissenschaftstheorie*. Berlin/Heidelberg: Springer, 6. Auflage.

Chan, Steve (1997): *In Search of Democratic Peace: Problems and Promise*, in: *Mershon International Studies Review* 41, S. 59–91.

Clauß, Günter und Heinz Ebner (1989): *Statistik für Soziologen, Pädagogen, Psychologen und Mediziner.* Thun, Frankfurt: Harri Deutsch, 6. Auflage.

Converse, Jean M. und Stanley Presser (1986): *Survey Questions. Handcrafting the Standardized Questionnaire*. Beverly Hills: Sage.

Diekmann, Andreas (2008): *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Reinbek bei Hamburg: Rowohlt, 19. Auflage.

Dillman, Don A. (1978): *Mail and Telephone Surveys. The Total Design Method*. New York u. a.: Wiley.

Ditton, Hartmut (1998): *Mehrebenenanalyse. Grundlagen und Anwendungen des Hierarchisch Linearen Modells*. Weinheim und München: Juventa.

Dorroch, Heinz (1994): *Meinungsmacher-Report. Wie Umfrageergebnisse entstehen*. Göttingen: Steidl.

Dovidio, J. F. und S. L. Gaertner (1986): *Prejudice, Discrimination, and Racism. Historical Trends and Contemporary Approaches*, in: J. F. Dovidio und S. L. Gaertner, Hrsg.: *Prejudice, Discrimination, and Racism.* Orlando: Academic Press, S. 1–34.

Druckman, James N. et al. (2006): *The Growth and Development of Experimental Research in Political Science*, in: *American Political Science Review* 100(4), S. 627–635.

Druwe, Ulrich (1994): *Studienführer Politikwissenschaft.* München: ars una, 2. Auflage.

Eagly, Alice H. und Shelly Chaiken (1993): *The Psychology of Attitudes.* Fort Worth u. a.: Harcourt Brace.

Elliot, Dave (1991): *Weighting for Non-Response. A Survey Researcher's Guide*. London: OPCS.

Engel, Uwe und Jost Reinecke (1994): *Panelanalyse. Grundlagen. Techniken. Beispiele*. Berlin: De Gruyter.

Fahrmeir, Ludwig et al. (2007): *Statistik. Der Weg zur Datenanalyse.* Berlin/Heidelberg: Springer, 6. Auflage.

Falter, Jürgen W. (1977): *Einmal mehr: Läßt sich das Konzept der Parteiidentifikation auf deutsche Verhältnisse übertragen? Theoretische, methodologische und empirische Probleme einer Validierung des Konstrukts Parteiidentifikation für die Bundesrepublik Deutschland*, in: *Politische Vierteljahresschrift* (2/3), S. 476–500.

Falter, Jürgen W. (1991): *Hitlers Wähler.* München: C. H. Beck.

Falter, Jürgen W. et al. (1983): *Arbeitslosigkeit und Nationalsozialismus. Eine empirische Analyse des Beitrags der Massenerwerbslosigkeit zu den Wahlerfolgen der NSDAP 1932 und 1933*, in: *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 35(3), S. 525–551.

Faulbaum, Frank, Marc Deutschmann und Martin Kleudgen (2003): *Computerunterstütztes Pretesting von CATI-Fragebögen: Das CAPTIQ-Verfahren*, in: *ZUMA-Nachrichten* 52, S. 20–34.

Flick, Uwe (2007): *Qualitative Sozialforschung.* Reinbek: Rowohlt.

Follmer, Richard und Menno Smid (1998): *Nichteingetragene Telefonnummern: Ergebnisse eines Methodentests*, in: Siegfried Gabler, Sabine Häder und Jürgen H.P. Hoffmeyer-Zlotnik, Hrsg.: *Telefonstichproben in Deutschland.* Opladen/Wiesbaden: Westdeutscher Verlag, S. 43–57.

Frey, Bruno S. und Hannelore Weck (1981): *Hat Arbeitslosigkeit den Aufstieg des Nationalsozialismus bewirkt?*, in: *Jahrbuch für Nationalökonomie und Statistik* 196, S. 1–31.

Frey, James H., Gerhard Kunz und Günther Lüschen (1990): *Telefonumfragen in der Sozialforschung. Methoden, Techniken, Befragungspraxis.* Opladen: Westdeutscher Verlag.

Friedrichs, Jürgen (1990): *Methoden empirischer Sozialforschung.* Opladen: Westdeutscher Verlag, 14. Auflage.

Friedrichs, Jürgen und Hartmut Lüdtke (1977): *Teilnehmende Beobachtung. Einführung in die sozialwissenschaftliche Feldforschung.* Weinheim: Beltz, 3. Auflage.

Früh, Werner (2007): *Inhaltsanalyse.* Konstanz: UVK, 6. Auflage.

Fuchs, Marek (1994): *Umfrageforschung mit Telefon und Computer.* Weinheim: Beltz.

Gabler, Siegfried und Sabine Häder (1998): *Probleme bei der Anwendung von RLD-Verfahren*, in: Siegfried Gabler, Sabine Häder und Jürgen H.P. Hoffmeyer-Zlotnik, Hrsg.: *Telefonstichproben in Deutschland.* Opladen/Wiesbaden: Westdeutscher Verlag, S. 58–68.

Gabler, Siegfried, Sabine Häder und Jürgen H.P. Hoffmeyer-Zlotnik, Hrsg. (1998): *Telefonstichproben in Deutschland.* Opladen: Westdeutscher Verlag.

Gabler, Siegfried, Jürgen H. P. Hoffmeyer-Zlotnik und Dagmar Krebs, Hrsg. (1994): *Gewichtung in der Umfragepraxis.* Opladen: Westdeutscher Verlag.

Gehring, Uwe W. und Jürgen R. Winkler (1997): *Parteiidentifikation, Kandidaten- und Issueorientierungen als Determinanten des Wahlverhaltens in Ost- und Westdeutschland*, in: Oscar W. Gabriel, Hrsg.: *Politische Orientierungen und Verhaltensweisen im vereinigten Deutschland.* Opladen: Leske + Budrich, S. 473–506.

GESIS, Hrsg. (2007): *ALLBUS. Datenhandbuch 2006. ZA-Nr. 4500.* Köln und Mannheim.

Grümer, Karl-Wilhelm (1974): *Beobachtung.* Stuttgart: Teubner.

Gschwend, Thomas (2005): *Ökologische Inferenz*, in: Joachim Behnke et al., Hrsg.: *Methoden der Politikwissenschaft. Neuere qualitative und quantitative Verfahren.* Baden-Baden: Nomos, S. 227–237.

Gschwend, Thomas und Marc Hooghe (2008): *Should I Stay or Should I Go? An Experimental Study on Voter Responses to Pre-electoral Coalitions*, in: *European Journal of Political Research* 47(5), S. 556 – 577.

Haisken-DeNew, John P. und Joachim R. Frick, Hrsg. (2005): *DTC. Desktop Companion to the German Socio-Economic Panel Study (GSOEP). Version 8.0 (Updated to Wave U).* Berlin: Deutsches Institut für Wirtschaftsforschung.

Hanefeld, Ute (1987): *Das Sozio-ökonomische Panel. Grundlagen und Konzeption.* Frankfurt/New York: Campus.

Harkness, Janet A., Fons van de Vijver und Peter Mohler, Hrsg. (2003): *Cross Cultural Survey Methods*. Hoboken, NJ: Wiley.

Häder, Sabine und Axel Glemser (2006): *Stichprobenziehung für Telefonumfragen in Deutschland*, in: Andreas Dieckmann, Hrsg.: *Methoden der Sozialforschung (Sonderheft 44 der Kölner Zeitschrift für Soziologie und Sozialpsychologie)*. Wiesbaden: VS Verlag, S. 148–171.

Häder, Sabine und Peter Lynn (2007): *How Representative can a Multi-Nation Survey be?*, in: Roger Jowell et al., Hrsg.: *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey*. Thousand Oaks u. a.: Sage, S. 33–52.

Hempel, Carl G. und Paul Oppenheim (1948): *Studies in the Logic of Explanation*, in: *Philosophy of Science* 15, S. 135–175.

Inglehart, Ronald (1977): *The Silent Revolution*. Princeton: Princeton University Press.

Jacob, Rüdiger und Willy H. Eirmbter (2000): *Allgemeine Bevölkerungsumfragen*. München/Wien: Oldenbourg.

Janetzko, Dietmar (1999): *Statistische Anwendungen im Internet. In Netzumgebungen Daten erheben, auswerten und praesentieren*. München: Addison-Wesley.

King, Gary S. (1997): *A Solution to the Ecological Inference Problem. Reconstructing Individual Behavior from Aggregate Data*. Princeton, New Jersey: Princeton University Press.

Koch, Achim (1991): *Zum Zusammenhang von Interviewermerkmalen und Ausschöpfungsquoten*, in: *ZUMA-Nachrichten* 15(28), S. 41–53.

Koch, Achim (1993): *Sozialer Wandel als Artefakt unterschiedlicher Ausschöpfung. Zum Einfluß von Veränderungen der Ausschöpfungsquote auf die Zeitreihen des ALLBUS*, in: *ZUMA-Nachrichten* 17, S. 83–113.

Koch, Achim (1995): *Gefälschte Interviews: Ergebnisse der Interviewerkontrolle beim* ALLBUS 1994, in: *ZUMA-Nachrichten* 19(36), S. 89–105.

Koch, Achim, Siegfried Gabler und Michael Braun (1994): *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS) 1994*, ZUMA-Arbeitsbericht 94/11, ZUMA, Mannheim.

Koch, Achim et al. (1999): *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS) 1998*, ZUMA-Arbeitsbericht 99/02, ZUMA, Mannheim.

Kohler, Ulrich (2006): *Schätzer für komplexe Stichproben*, in: Joachim Behnke et al., Hrsg.: *Methoden der Politikwissenschaft. Neuere qualitative und quantitative Verfahren*. Baden-Baden: Nomos, S. 309–320.

Kohler, Ulrich und Frauke Kreuter (2008): *Datenanalyse mit Stata. Allgemeine Konzepte der Datenanalyse und ihre praktische Anwendung*. München/Wien: Oldenbourg, 3. Auflage.

Krämer, Walter (1991): *So lügt man mit Statistik*. Frankfurt/New York: Campus.

Krämer, Walter (1994): *So überzeugt man mit Statistik*. Frankfurt/New York: Campus.

Krügener, Sonja (2006): *Registergestützter Zensus – Aktueller Stand und Entwicklungsperspektiven*, in: Frank Faulbaum und Christof Wolf, Hrsg.: *Stichprobenqualität in Bevölkerungsumfragen*. Bonn: IZ, S. 55–61.

Kromrey, Helmut (2006): *Empirische Sozialforschung*. Stuttgart: Lucius & Lucius, 11. Auflage.

Kühnel, Steffen und Dagmar Krebs (2007): *Statistik für die Sozialwissenschaften. Grundlagen, Methoden, Anwendungen*. Reinbek bei Hamburg: Rowohlt, 4. Auflage.

Kurz, Karin, Peter Prüfer und Margit Rexroth (1999): *Zur Validität von Fragen in standardisierten Erhebungen. Ergebnisse des Einsatzes eines kognitiven Pretestinterviews*, in: *ZUMA-Nachrichten* 44, S. 83–108.

Lakatos, Imre (1974): *Falsifikation und die Methodologie wissenschaftlicher Forschungsprogramme*, in: Imre Lakatos und Alan Musgrave, Hrsg.: *Kritik und Erkenntnisfortschritt*. Braunschweig: Vieweg, S. 89–189.

Lazarsfeld, Paul F., Bernard Berelson und Hazel Gaudet (1968): *The People's Choice. How the Voter Makes Up His Mind in a Presidential Campaign*. New York und London: Columbia University Press, 3. Auflage.

Lazarsfeld, Paul F. und Herbert Menzel (1972): *Group Characteristics and Their Interrelations*, in: Paul F. Lazarsfeld, Ann K. Pasanella und Morris Rosenberg, Hrsg.: *Continuities in the Language of Social Research*. New York: The Free Press, S. 225–237.

Lechert, Yvonne, Paul Schroedter und Paul Lüttinger (2006): *Die Umsetzung der Bildungsklassifikation CASMIN für die Volkszählung 1970, die Mikrozensus- Zusatzerhebung 1971 und die Mikrozensen 1976-2004. ZUMA-Methodenbericht 2006/12*. Mannheim: ZUMA.

Levy, Paul und Stanley Lemeshow (1991): *Sampling of Populations: Methods and Applications*. New York u. a.: Wiley, 2. Auflage.

Little, Roderick J. A. und Donald B. Rubin (2002): *Statistical Analysis With Missing Data*. New York: Wiley, 2. Auflage.

Long, J. Scott (1997): *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks u. a.: Sage.

Mayntz, Renate, Kurt Holm und Peter Hübner (1978): *Einführung in die Methoden der empirischen Soziologie*. Opladen: Westdeutscher Verlag, 5. Auflage.

Mayring, Philipp (2007): *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Weinheim: Deutscher Studienverlag, 9. Auflage.

McIver, John P. und Edward G. Carmines (1982): *Unidimensional Scaling*. London: Sage, 2. Auflage.

Merten, Klaus (1995): *Inhaltsanalyse. Einführung in Theorie, Methode und Praxis*. Opladen: Westdeutscher Verlag, 2. Auflage.

Mletzko, Matthias und Cornelia Weins (1999): *Polizei und Fremdenfeindlichkeit - Ergebnisse einer Befragung in einer westdeutschen Polizeidirektion.*, in: *Monatsschrift für Kriminologie und Strafrechtsreform* 82(2), S. 77–93.

Müller, Walter und Yossi Shavit (1998): *Bildung und Beruf im institutionellen Kontext. Eine vergleichende Studie in 13 Ländern*, in: *Zeitschrift für Erziehungswissenschaft* 1, S. 501–533.

Mohler, Peter Ph. und Wolfgang Bandilla, Hrsg. (1992): *Blickpunkt Gesellschaft 2. Einstellungen und Verhalten der Bundesbürger in Ost und West*. Opladen: Westdeutscher Verlag.

Müller, Walter et al., Hrsg. (1990): *Blickpunkt Gesellschaft. Einstellungen und Verhalten der Bundesbürger*. Opladen: Westdeutscher Verlag.

Noelle-Neumann, Elisabeth und Thomas Petersen (1996): *Alle, nicht jeder. Einführung in die Methoden der Demoskopie*. München: dtv.

Opp, Karl-Dieter (2005): *Methodologie der Sozialwissenschaften*. Wiesbaden: VS Verlag, 6. Auflage.

Orth, Bernhard (1974): *Einführung in die Theorie des Messens*. Stuttgart: Kohlhammer.

Pappi, Franz U. (1977): *Aggregatdatenanalyse*, in: Jürgen van Koolwijk und Maria Wieken-Mayser, Hrsg.: *Techniken der empirischen Sozialforschung. Band 7: Datenanalyse*. München/Wien: Oldenbourg, S. 78–110.

Patzelt, Werner J. (2007): *Einführung in die Politikwissenschaft*. Passau: Wissenschaftsverlag Rothe, 6. Auflage.

Popper, Karl R. (1971): *Die Logik der Forschung*. Tübingen: Mohr.

Porst, Rolf (1985): *Praxis der Umfrageforschung. Erhebung und Auswertung sozialwissenschaftlicher Umfragedaten*. Stuttgart: Teubner.

Poser, Hans (2001): *Wissenschaftstheorie. Eine philosophische Einführung*. Stuttgart: Reclam.

Prim, Rolf und Heribert Tilmann (1997): *Grundlagen einer kritisch-rationalen Sozialwissenschaft*. Wiesbaden: Quelle und Meyer, 7. Auflage.

Raschke, Joachim (1991): *Die Parteitage der Grünen*, in: *Aus Politik und Zeitgeschichte* 41(B11-12), S. 46–54.

Reinecke, Jost (1991): *Interviewer- und Befragtenverhalten. Theoretische Ansätze und methodische Konzepte*. Opladen: Westdeutscher Verlag.

Reinecke, Jost (2005): *Strukturgleichungsmodelle in den Sozialwissenschaften*. München: Oldenbourg.

Ritsert, Jürgen (2003): *Einführung in die Logik der Sozialwissenschaften*. Münster: Westfälisches Dampfboot, 2. Auflage.

Robinson, William S. (1950): *Ecological Correlations and Behavior of Individuals*, in: *American Sociological Review* 15, S. 351–357.

Rohwer, Götz und Ulrich Pötter (2002): *Wahrscheinlichkeit. Begriff und Rhetorik in der Sozialforschung.* Weinheim und München: Juventa.

Rost, Jürgen (2004): *Lehrbuch Testtheorie und Testkonstruktion.* Bern u. a.: Huber, 2. Auflage.

Rucht, Dieter, Peter Hocke und Dieter Oremus (1995): *Quantitative Inhaltsanalyse: Warum, wo, wann und wie wurde in der Bundesrepublik protestiert?*, in: Ulrich von Alemann, Hrsg.: *Politikwissenschaftliche Methoden.* Opladen: Westdeutscher Verlag, S. 261–291.

Sachs, Lothar (2006): *Angewandte Statistik.* Berlin/Heidelberg: Springer, 12. Auflage.

Sarris, Viktor (1999): *Einführung in die experimentelle Psychologie.* Lengerich u. a.: Pabst.

Schafer, J. L. und J. W. Graham (2002): *Missing Data: Our View of the State of the Art*, in: *Psychological Methods* 7(2), S. 147–177.

Scheaffer, Richard L., William Mendenhall und Lyman Ott (1996): *Elementary Survey Sampling.* Belmont, Ca.: Wadsworth, 5. Auflage.

Schmitt-Beck, Rüdiger, Stefan Weick und Bernhard Christoph (2006): *Shaky Attachments: Individual-level Stability and Change of Partisanship Among West German Voters, 1984-2001*, in: *European Journal of Political Research* 45(4), S. 581 – 608.

Schnell, Rainer (1991): *Der Einfluß gefälschter Interviews auf Survey-Ergebnisse*, in: *Zeitschrift für Soziologie* 20(1), S. 25–35.

Schnell, Rainer, Paul B. Hill und Elke Esser (2008): *Methoden der empirischen Sozialforschung.* München/Wien: Oldenbourg, 8. Auflage.

Schoen, Harald (2000): *Den Wechselwählern auf der Spur: Recall- und Paneldaten im Vergleich*, in: Jan van Deth, Hans Rattinger und Edeltraud Roller, Hrsg.: *Die Republik auf dem Weg zur Normalität? Wahlverhalten und politische Einstellungen nach acht Jahren Einheit.* Opladen: Leske + Budrich, S. 199–226.

Schuman, Howard und Stanley Presser (1996): *Questions and Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context (Nachdruck)*. Thousand Oaks: Sage.

Schwarz, Norbert et al. (1985): *Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments*, in: *Public Opinion Quarterly* 49, S. 388–395.

Snijders, Tom A. B. und Roel J. Bosker (1999): *Multilevel Analysis*. London u. a.: Sage.

Stevens, Stanley S. (1946): *On the Theory of Scales of Measurements*, in: *Science* 103, S. 677–680.

Sudman, Seymour (1982): *Asking Questions. A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass.

Teusch, Ulrich und Martin Kahl (2001): *Ein Theorem mit Verfallsdatum? Der 'Demokratische Frieden' im Kontext der Globalisierung*, in: *Zeitschrift für Internationale Beziehungen* 8(2), S. 287–320.

Tuckey, John W. (1977): *Exploratory Data Analysis*. Reading: Addison Wesley.

Wagner, Gert G. (2006): *Die Macht der Zahlen*, in: *360 Grad. Fachmagazin für das Management im öffentlichen Sektor* (5), S. 14–15.

Wallace, Walter L. (1971): *The Logic of Science in Sociology*. Chicago/New York: Aldine-Atherton.

Wasmer, Martina et al. (1996): *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 1996*, ZUMA-Arbeitsbericht 96/08, ZUMA, Mannheim.

Wasmer, Martina, Evi Scholz und Michael Blohm (2007): *Konzeption und Durchführung der „Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) 2006*, ZUMA-Methodenbericht 2007/09, ZUMA, Mannheim.

Weber, Max (1980): *Wirtschaft und Gesellschaft. Grundriß der verstehenden Soziologie*. Tübingen: Mohr, 5. Auflage.

Weber, Robert P. (1990): *Basic Content Analysis*. Newbury Park: Sage, 2. Auflage.

Widmaier, Ulrich (1997): *Vergleichende Aggregatdatenanalyse*, in: Dirk Berg-Schlosser und Ferdinand Müller-Rommel, Hrsg.: *Vergleichende Politikwissenschaft*. Opladen: Leske + Budrich, 3. Auflage, S. 103–118.

Winkler, Jürgen (1995): *Sozialstruktur, politische Traditionen und Liberalismus. Eine empirische Längsschnittstudie zur Wahlentwicklung in Deutschland 1871–1933*. Opladen: Westdeutscher Verlag.

Wooldridge, Jeffrey M. (2006): *Introductory Econometrics. A Modern Approach*. Cincinnati, Ohio: South-Western College Publ.

Zentralarchiv für empirische Sozialforschung (1999): *Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. ALLBUS 1998. Codebuch ZA-Nr. 3000*. Zentralarchiv für Empirische Sozialforschung an der Universität zu Köln: Köln.

Register

U. W. Gehring, C. Weins, *Grundkurs Statistik für Politologen und Soziologen*,
DOI 10.1007/978-3-531-91879-2,